Proceedings

FONETIK 2018

The XXXth Swedish Phonetics Conference
Gothenburg, June 7–8

音声学

Proceedings FONETIK 2018

The XXXth Swedish Phonetics Conference,
held at University of Gothenburg, June 7–8, 2018

Edited by Åsa Abelin and Yasuko Nagano-Madsen
Department of Philosophy, Linguistics and Theory of Science University and Department of Languages and Literatures

University of Gothenburg
Box 200, SE 405 30 Gothenburg

Printed by Reprocentralen, Humanisten, University of Gothenburg.

Cover photo by Monica Havström

Preface

The XXX on the front page of this volume is not a place holder for an exact Roman numeral, but this year's conference is in fact the thirtieth Swedish Phonetics Conference. Only twice have we not had the annual phonetics meeting, once in 1995 when ICPhS was held in Stockholm and last year in 2017 when Interspeech was held, also in Stockholm.

We hope that this year's meeting in Gothenburg will be as interesting and enjoyable as all earlier Swedish phonetics meetings, with both old and new members of our community from the Nordic countries and all around the world. We thank Fonetikstiftelsen for generous financial support.

Gothenburg, June 2018

Åsa Abelin and Yasuko Nagano-Madsen

Previous Swedish Phonetics Conferences (from 1986)

https://www.ling.su.se/fonetik-2014/tidigare-konferenser

Contents

Phonaesthemes in older Swedish dialects – the clusters *fn-*, *gn-*, *skv-*, *pj-* and *kn-*

Åsa Abelin
Department of philosophy, linguistics and theory of science, University of Gothenburg

Abstract

The purpose of this study is to find similarities and differences of onomatopoeic and sound symbolic words in older Swedish dialects, the so called genuine dialects, in comparison with contemporary usage. The comparison is both qualitative and quantitative. By making an inventory of older dialect data bases, we can establish onomatopoeia and sound symbolism in a way which is complementary to analyses already made of sound symbolism in standard Swedish. The questions are whether phonaesthemes have disappeared or been created new, whether words have disappeared or been created new, and whether the proportions of sound symbolic words and phonaesthemes are the same or not.

The study has excerped words from the central Swedish dialect dictionary written by Rietz (1862-1867), which contains words from the middle of the 19th century. The method follows the procedure used for the establishment of phonaesthemes (consonant combinations with certain meanings) in present day standard Swedish (Abelin, 1999). The initial consonant clusters and all meanings which were found to be sound symbolic will be the basis for comparison. The percentually most common phonaesthemes in standard Swedish, fn-, gn-, skv-, pj-, kn-, were studied in the older dialects.

The results show that 1) some phonaesthemes have disappeared, but no new phonasesthemes have appeared 2) many sound symbolic dialectal words have disappeared 3) distributions of meanings for clusters have changed 4) the order of the five most common sound symbolic clusters has not changed.

Introduction

Phonaesthemes are part of sound symbolism in language and can briefly be described as consonant clusters which have in common a certain element of meaning, and concern both onomatopoeic and sound symbolic meanings. The difference between phonasthemes and morphemes can be seen in the following way. Phonaesthemes do not involve whole syllables with a vowel, and regular morphemes can sometimes include one or more phonaesthemes. Phonaesthemes can be productive at a slow pace and are often used (or excluded) in the creation of efficient brand names (Abelin, 2015). Sound symbolism has recently received a lot of attention e.g. in child language research. For a recent overview of the subject, see Svantesson (2017). In Abelin (1999) the inventory of phonaesthemes in present day standard Swedish was analysed and described. This study will make a comparison between phonaesthemes in present day Swedish and in older Swedish dialects. The focus is on which words and which phonaesthemes have disappeared or been created new. Some examples of new creations in child language, advertising language, literature or just in general (including borrowings) are: *bling (shining detail), blippa (blip), dongle (dongle), pjämmel (pejorative for lamentation), knasprig (about sound and texture of crisps), fläbbig (pejorative about wrinkled face), huckra (type of sound from a car engine), skrövla (wrinkle), drissla (spread in small chunks), spritsig (about sour and tingly taste).* Some of these have established themselves in the lexicon and others are temporary creations. They all seem to follow a general pattern however, and we seem to understand them.

Phonaesthemes

The meaning of phonaesthemes is here defined as a string of initial consonants which have in common a certain element of meaning. The meaning is related to one or more of the senses hearing, sight or touch, as in descriptions of impressions of sound (onomatopoeia), light,

movement, form or surface structure. The meaning can also be related to emotions or attitudes. The point is that there is a non-arbitrary connection – iconic or indexical – between expression and meaning.

Rietz Swedish dialect lexicon

The words for the dialect lexicon were collected by Rietz himself travelling all over Sweden. He also had the help from 26 assistants (often clergymen). Furthermore he used dialectal word lists collected by other persons, which he controlled with local people. He was truly interested in the language of the people and he did not shy away from offensive words. The dictionary was published between 1862 and 1867 and the material was collected in the years before that, meaning that the words were collected over 150 years ago.

The organization of the lemmas and lexemes in the dictionary is somewhat difficult to follow and for this reason Abrahamsson (Rietz & Abrahmsson, 1955) made a separate register of the headwords.

Research questions

- Do the same phonaesthemes exist in present day Swedish and the older dialects?
- Do the same words exist in the older dialects, and in modern standard Swedish?
- Did the same consonant cluster exist, but with other meanings, i.e. other phonaesthemes?
- Are the lexical proportions the same in present day Swedish and the older dialects?

Method

The dialect lexicon by Rietz was excerpted in order to make a comparison with the analysis of phonaesthemes in SAOL 10 (Abelin, 1999). The percentually most common phonesthemes in Abelin (1999) were *fn-, gn-, skv-, pj-* and *kn-*. *Fn-* is typically pejorative or onomatopoeic, *gn-* imitates sound, speech, light, and is pejorative, *skv-* is onomatopoeic, expecially of water sounds, *pj-* is pejorative and *kn-* imitates round form, is onomatopoeic and pejorative. As an example of the analysis, the cluster *fn-* has 10 sound symbolic root morphemes out of a total of 10 root morphemes beginning with *fn-*; therefore the percent sound symbolic root morphemes is 100%. For *gn-* the percent is 92%, for *skv-* 90%, for *pj-* 76%, and for *kn-* 67%. In reaction time experiments (Abelin, 2012) exactly these clusters have also shown the largest psychological reality. Based on the frequencies in Abelin (1999, 2012), all words beginning with these percentually five most common initial consonant clusters were excerpted from Rietz dialect lexicon. In the following both the absolute numbers and the percentages in Abelin (1999) and Rietz (1862-1867) will be compared.

In order to count the number of dialectal words on a certain consonant cluster I have used the register, created by Abrahamsson (Rietz & Abrahamsson, 1955). In the following "words" are used interchangeably with "root morphemes".

A word of caution: The analysis of SAOL 10 (Abelin, 1999) is built on updated material for standard language in a dictionary which has been published in several editions for a long period of time, while the dialect dictionary of Rietz builds on spoken material collected more than 150 years ago. Therefore, the differences between the two dictionaries do not necessarily have to do with older versus newer language use, but could also depend on standard language versus dialects. In light of these differences, eventual similarities between the two lexica are very interesting as well.

Results

The results are first presented for each cluster. Absolute frequencies of sound symbolic words are first presented. For *fn-* SAOL 10 has 10 root morphemes, and Rietz has 53 sound symbolic root morphemes. Many of the dialectal words are no longer in use, and some words probably had a slightly different meaning than today. What seems to be new discovered phonaesthemes (i.e. phonaesthemes which have disappeared) are *fn-* "roughness" as in *fnarre*, "form" as in *fnöl*, *fnaga*, "angry feeling" as in *fnöst* and "hurry up" as in *fnalla.* Words and phonaesthemes have disappeared. In some cases the meaning of a word has changed, as in *fnatta* meaning "walk like a child". Many of the disappeared words had to do with lice, itching, bumps and abscesses which are not as common among the population today.

Gn- has 20 sound symbolic root morphemes in SAOL, and 97 sound symbolic root morphemes in Rietz. A disappeared phonaestheme is "stinginess" as in *gnuver* "stingy person". Many words have disappeared and words have changed

their meanings, as *gnola* which meant "pray incessantly like a child".

The cluster *skv-* had 9 sound symbolic root morphemes in SAOL and 64 in Rietz. *Pj-* had 6 sound symbolic root morphemes in SAOL and 51 in Rietz. *Kn-* had 59 sound symbolic root morphemes in SAOL and 249 in Rietz. Cf. also percentages in Figure 3.

For none of these five clusters there are newly created phonaesthemes in present day Swedish.

The clusters presented as phonaesthemes

In the following diagrams in Figure 1 and 2 we can see the absolute numbers of words for different phonaesthemes. The *kn-* phonaesthemes are excluded here because there are so many instances of these. The scales are not identical in the two diagrams due to the fact that Rietz dictionary contains many more sound symbolic words.

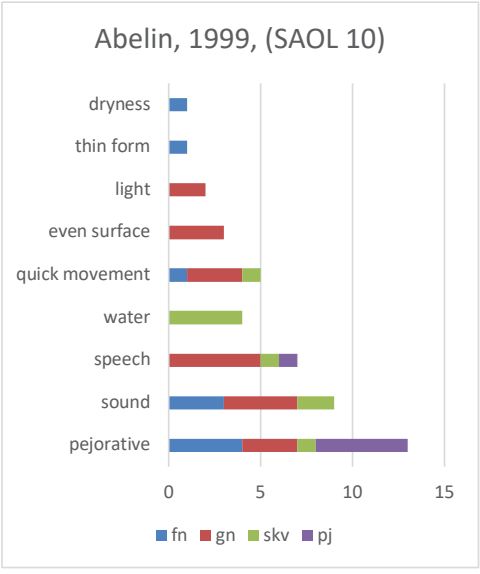

Figure 1. Absolute frequencies of root morphemes for different phonaesthemes according to analyses of SAOL 10 (Abelin, 1999). Scale 1–15.

As an example, the meaning "pejorative" is found for all the four clusters, in both corpora, but in Rietz there are many more occurences of pejorative words. As a contrast, "laughter" is only found in Rietz, and only for the clusters *fn-* and *gn-*.

Figure 1 will be discussed in conjunction with Figure 2 below.

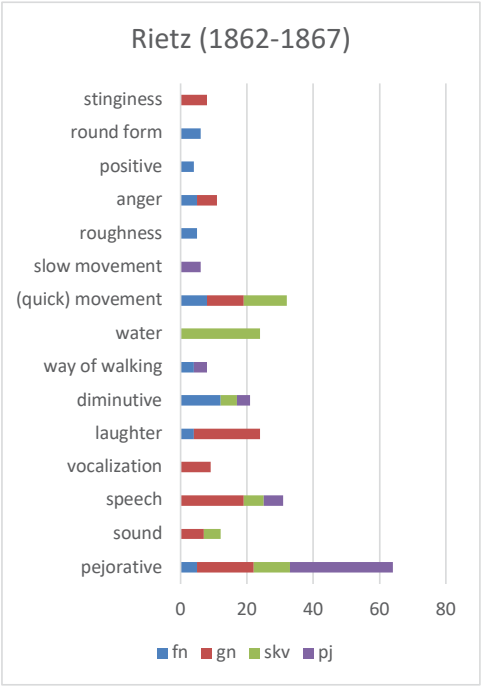

Figure 2. Absolute frequencies of root morphemes for different phonaesthemes according to analyses of Rietz (1862–1867). Scale 1–80.

Figures 1 and 2 show a number of similarities and differences. The most striking result is that in the dialects there were both more sound symbolic meanings involved (9 in Abelin, and 15 in Rietz), and more sound symbolic words (e.g. more than 60 words for pejoratives in Rietz, but only 13 in Abelin). Furthermore, the category "diminutive" was common in the dialects, expecially for *fn-*, but is not occurring today for these clusters. *Fn-* seems to have become more "pejorative" than before. *Gn-* used to carry the meanings of "stinginess" and "laughter", but today it is more associated with "light", and "smooth surface". *Skv-* was earlier more associated with "movement" and has become more associated with "sound", "speech" and "pejorative". Finally, *pj-* used to carry the meaning of "slow movement".

Proportion of sound symbolic root morphemes in relation to total number of root morphemes

The proportion of sound symbolic root morphemes in relation to number of root morphemes for each of the five consonant clusters was analysed. It is difficult to compare the percentages between two lexica which are structured so differently, but measuring the percent sound symbolic words of the clusters gave the following result, see Figure 3. The differences between each of the consonant clusters are small, but nevertheless interesting. There is a positive correlation between percent sound symbolic words for the five clusters in Rietz and SAOL, r=0.924. An interesting question is the reason for a higher percent sound symbolic words in SAOL; could it be the case that these sound symbolic clusters are becoming more marked and thus less useful for arbitrary words?

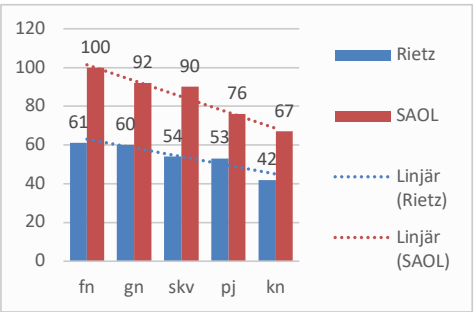

Figure 3. Percent sound symbolic words in relation to number of words for each of the five consonant clusters, in Rietz and SAOL.

In the analyses of phonaesthemes in both lexica the ranking order for the highest proportion of sound symbolic words to the lowest proportion of sound symbolic words is the same: fn > gn > skv > pj > kn. This ranking order is the same in both older Swedish dialects and present day standardised Swedish.

Discussion

The comparison of the two dictionaries is possible even though they are structured differently. Both more in-depth studies and more extended studies of all phonesthemes would contribute to a more thorough description of phonaesthemes in the older Swedish dialects and of language change. Apart from the contributions to linguistic analyses, the old dialectal words and phonaesthemes bear testimony of older living conditions. They give a picture of life in misery and deprivation, negative attitudes to people who are lazy, slow, immoral, stingy or angry, and for watery grain and bad harvests. But there are also many words for talking and laughter, and for (funny) ways of walking. Another intriguing subject for further studies of newer and older sound symbolic Swedish words is the comparison with sound symbolism in Meänkieli in order to examine possible contributions from older Swedish and Finnish.

Conclusions

- Some of the phonaesthemes in the dialects are not existing in standard Swedish today. No new phonaesthemes have appeared.
- Many of the sound symbolic dialectal words are not existing in standard Swedish today.
- The distribution of the meanings over the consonant clusters are partly the same, partly different today.
- The sound symbolic size ratio between the five consonant clusters show tendencies to be the same in the older Swedish dialects and in standard Swedish of today: fn > gn > skv > pj > kn.

References

Abelin Å (1999). *Studies in Sound Symbolism* Gothenburg Monographs in Linguistics 17. Göteborg.

Abelin Å (2012). Relative frequency and semantic relations as organizing principles for the psychological reality of phonaesthemes, *Selected articles from UK-CLA meetings, vol 1*, 128–145.

Abelin Å (2015). Phonaesthemes and sound symbolism in Swedish brand names, *Ampersand vol. 2,* 19–29.

Rietz JE (1862-1867/1962). *Svenskt dialektlexikon; ordbok öfver svenska allmogespråket,* CWK Gleerups förlag, Lund.

Rietz JE & Abrahamsson E (1955). *Svenskt dialektlexikon: ordbok öfver svenska allmogespråket. 3, Register och rättelser.* Uppsala: Lundequistska bokh.

Svantesson J (2017). Sound symbolism: the role of word sound in meaning. *WIREs Cogn Sci*, 8:e1441.

Svenska akademiens ordlista över svenska språket (1973). 10:e uppl. (SAOL 10), Stockholm: P. A. Norstedt & Söners förlag.

Accentual falls and rises vary as a function of accompanying head and eyebrow movements

Gilbert Ambrazaitis[1] and David House[2]
[1]Department of Swedish, Linnaeus University, Växjö, Sweden
[2]Department of Speech, Music and Hearing, KTH, Stockholm

Abstract

In this study we examine prosodic prominence from a multimodal perspective. Our research question is whether the phonetic realization of accentual falls and rises in Swedish complex pitch accents varies as a function of accompanying head and eyebrow movements. The study is based on audio and video data from 60 brief news readings from Swedish Television (SVT Rapport), comprising 1936 words in total, or about 12 minutes of speech from five news anchors (two female, three male). The results suggest a tendency for a cumulative relation of verbal and visual prominence cues: the more visual cues accompanying, the higher the pitch peaks and the larger the rises and falls.

Introduction

Previous research on co-speech gestures and audio-visual prosody strongly suggests that prosodic prominence is indeed an audio-visual, or multimodal, phenomenon: Pitch accents (verbal prominence cues) are frequently accompanied by movements of the hands, the head and certain facial areas (visual cues), also referred to as beat gestures (e.g., Kendon, 1980, McClave, 2000).

It has, moreover, been shown that visual and verbal prominence cues may co-occur in various constellations (Swerts & Krahmer, 2010; Loehr, 2012; Ambrazaitis & House, 2017) and that beat gestures are more likely to occur with *perceptually strong* accents than with *weak* ones: Swerts and Krahmer (2010) found in their study of Dutch news readings that the more accented a word was on an auditory scale, the more likely the word was to also be accompanied by a head movement, an eyebrow movement or both. Hence, we might predict a *cumulative relation* of verbal and visual prominence cues, i.e. a positive correlation between the acoustic strength of a pitch accent (e.g. in terms of segmental durations, F0 peak height or F0 ranges) and accompanying beat gestures.

In this study, we test this prediction for the special case of complex pitch accents in a corpus of Stockholm Swedish news readings. Our research question is whether the phonetic realization of accentual rises and falls in such accents varies as a function of accompanying beat gestures by the head and the eyebrows.

Swedish makes use of pitch contrasts at the lexical level, distinguishing between two so-called word accents (Accent 1, Accent 2). Orthogonal to this word accent contrast, many varieties of Swedish, including the Stockholm dialect studied here, distinguish between two pitch-related phonological prominence levels, where the higher level has been commonly referred to as a *focal* accent or more recently as a *big* accent (Myrberg & Riad, 2015), as opposed to the *non-focal* or *small* accent (Myrberg & Riad, 2015). Crucially, the distinction between Accent 1 and Accent 2 is encoded at both levels. According to Bruce's (1977) seminal analysis, the big accent can be conceived of as a complex pitch accent composed of the tonal configuration for the word accent (Accent 1 or 2) and a following high tone H- (the *sentence accent* in his analysis, cf. Bruce, 1977) which is realized as a rise in pitch from the accentual L (HL* in Accent 1, H*L in Accent 2). Although the details of tonal representation of Swedish word accents, as well as the question of the lexicality of tones involved, is much debated (e.g. Bruce, 1977; Myrberg & Riad 2015; Wetterlin et al., 2007), there is a certain consensus on the compositional nature of big accents, as well as the assumption that the tonal components of a big accent relate to different prominence levels and different domains in the prosodic hierarchy of Swedish (Myrberg & Riad, 2015). In this brief report, we simplify the debate by referring to the tones defining the word accents as *accentual*, and the subsequent rise (the H- tone) as the *sentence accent rise*, (cf. Bruce, 1977).

The choice of complex (or big) pitch accents in Stockholm Swedish as an object of study thus introduces a prosodic-phonological dimension to our research question: Do we find a cumulative relation between the occurrence of head and eyebrow movements and (a) the fall at the accentual level, (b) the rise at the sentence level, or (c) both components of a complex pitch accent in Swedish? Answering this question would add to our general understanding of gesture-speech integration, and more specifically of the interaction of visual and verbal prominence cues.

Method

The present study is based on audio and video data from 60 brief news readings from Swedish Television (SVT Rapport), comprising 1936 words in total, or about 12 minutes of speech from five news anchors (two female, three male). The material was transcribed, segmented at the word level, and annotated for big accents (henceforth, BA), head beats (HB) and eyebrow beats (EB) using a combination of ELAN (Sloetjes & Wittenburg, 2008) and Praat (Boersma, 2001). In a first step of annotation, the presence of BA, HB and EB was judged upon on a word-basis. About half of the materials (30 files) were annotated by three labelers independently of each other. Inter-rater reliability was tested using Fleiss' κ (Fleiss, 1971), and turned out fair to good (BA: $\kappa = 0.77$; HB: $\kappa = 0.69$; EB: $\kappa = 0.72$). For the purpose of this study, the analysis focuses on three conditions:

- (i) words with a BA only (i.e. without a beat gesture: 276 tokens in our material),
- (ii) words with BA co-occurring with a HB (BA+HB: 178 tokens)
- (iii) words with BA co-occurring with HB and EB (BA+HB+EB: 73 tokens)

In a second step of annotation, tonal targets were labelled for all 527 tokens of interest: (H+)L* H- in case of Accent 1 and H*+L H- in case of Accent 2 (where H- is the sentence accent tone). In addition, as a baseline condition, tonal targets were labelled for a random selection of 102 non-focally accented words (small accents: 52 Accent 1, 50 Accent 2). We treat conditions (i-iii) above with the baseline condition added as a four-level independent variable (or fixed factor) in our analysis (see Results section) and refer to this variable as *multimodal prominence cluster* (henceforth, **MMP**).

Based on these tonal annotations, seven measures were derived to capture different aspects of the phonetic realization of the accentual fall (HL* or H*L respectively), and the sentence accent rise (H-):

- **I** – absolute peak height of the accentual fall (HL*/H*L) in Hz
- **II** – absolute peak height of the sentence accent rise (H-) in Hz
- **III** – range of the accentual fall (HL*/H*L) in semitones
- **IV** – range of the sentence accent rise (H-) in semitones
- **V** – highest peak in word (= either HL*/H*L or H-)
- **VI** – largest range in word (= either accentual fall or sentence accent rise)
- **VII** – the difference between H- and the preceding HL*/H*L in semitones.

Results and discussion

The results reveal slight effects of the factor **MMP** (cf. above) on measures **I-VI**. Figures 1-3 display the results for measures **III**, **IV**, and **VI** as an example; the corresponding illustrations for measures **I, II** and **V** provide a similar picture. No effect of MMP is observed for measure **VII**.

These results suggest a tendency for a cumulative relation of verbal and visual prominence cues: i.e., the more visual cues accompanying, the higher the pitch peaks and the larger the rises and falls. These effects are strongest for the combined measures **V** and **VI** (cf. Fig 3). All dependent variables were analyzed by means of linear mixed effects models assuming three fixed factors:

- *MMP* (cf. above)
- *speaker sex*
- *word accent* (Accent 1, Accent 2),

In addition, *speaker* was included as a random effect.

According to a likelihood ratio test for each of the seven dependent variables (i.e. measures **I-VII**), the effect of MMP was significant for measures **V** (p=.015*) and **VI** (p=.042*) suggesting a correlation of verbal and visual prominence cues that is reflected in both components of complex pitch accents in Stockholm Swedish – the accentual fall and the sentence accent rise.

The resuls from this study thus lend *acoustic* support for the *perception*-based prediction of a *cumulative relation* of verbal and visual prominence cues, a prediction we derived from Swerts and Krahmer's (2010) results based on auditory ratings for Dutch.

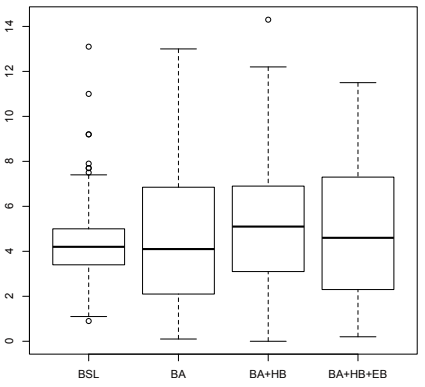

Figure 1. Boxplot for measure III – range of the accentual fall (HL/H*L) in semitones as a function of the multimodal prominence cluster (MMP); BSL = small accents; BA= big accents; HB=head heat; EB = eyebrow beat.*

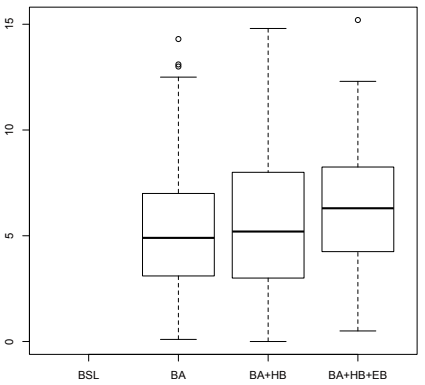

Figure 2. Boxplot for measure IV – range of the sentence accent rise (H-) in semitones as a function of MMP (cf. Fig. 1 for explanations).

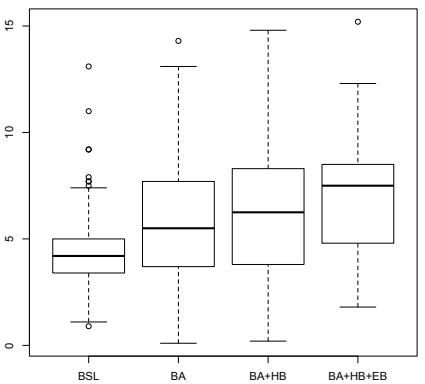

Figure 3. Boxplot for measure VI – largest range in word in semitones as a function of MMP (cf. Fig. 1 for explanations).

At the conference, the results are further discussed both in the light of a proposed outline of a model of multimodal prominence production, and with reference to prosodic domains assumed in Swedish phonology.

Acknowledgements

We retrieved materials from the National Library of Sweden and received permissions from Swedish Television. We also thank our research assistants Malin Svensson Lundmark, Anneliese Kelterer, and Otto Ewald for assistance with data processing and annotations. This work was supported by the Marcus and Amalia Wallenberg Foundation [MAW 2012.01.03] and the Swedish Research Council [VR 2017-02140].

References

Ambrazaitis G, House D (2017). Multimodal prominences: Exploring the patterning and usage of focal pitch accents, head beats and eyebrow beats in Swedish television news readings. *Speech Communication*, 95: 100-113.

Boersma P (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5: 341-345.

Bruce, G (1977). *Swedish Word Accents in Sentence Perspective*. Lund: Glerup.

Fleiss J (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76: 378-382.

Kendon A (1980). Gesticulation and speech: Two aspects of the process of utterance. In: M R Key, ed, *The relationship of verbal and nonverbal communication.*The Hague: Mouton, 207-227.

Loehr D (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology. Journal of the Association for Laboratory Phonology,* 3: 71-889.

McClave E (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics,* 32: 855-878.

Myrberg S, Riad T (2015). The prosodic hierarchy of Swedish. *Nordic Journal of Linguistics*, 23: 115-147.

Sloetjes H, Wittenburg P (2008). Annotation by category - ELAN and ISO DCR. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*.

Swerts M, Krahmer E (2010). Visual prosody of newsreaders: Effects of information structure, emotional content and intended audience on facial expressions. *Journal of Phonetics,* 38: 197-206.

Wetterlin A, Jönsson-Steiner E, Lahiri A (2007). Tones and loans in the history of Scandinavian. In: T Riad, C Gussenhoven, eds, *Tones and Tunes Volume 1*. Berlin; New York: Mouton de Gruyter.

Språkbanken Tal: A national research infrastructure for speech technology

Jens Edlund and David House
Dept of Speech, Music and Hearing/Språkbanken Tal/CLARIN Speech/Swe-Clarin
KTH Royal Institute of Technology

Abstract

In 2018, the national research infrastructure Nationella Språkbanken (NS) received funding from the Swedish Research Council. The funding of NS secures and extends Språkbanken in Gothenburg, which has provided resources for text based language research since the early 70s. This announcement introduces Språkbanken Tal, a branch of Nationella Språkbanken which is hosted by KTH and which will provide a much-needed infrastructure for speech science and speech technology.

Introduction

In 2018, *Språkbanken* in Gothenburg became *Nationella Språkbanken* (NS), through funding in the Swedish Research Council's programme for national research infrastructures. In addition to securing the continued operation of the original Språkbanken (now *Språkbanken Text*), Nationella Språkbanken adds two new branches: *Språkbanken Sam* (Eng. "Society"), operated by the Swedish Language Council at the Institute for Language and Folklore (ISOF), which supports research on the languages, dialects and other parts of the intangible cultural heritage in Sweden, and *Språkbanken Tal* (Eng. "Speech"), which caters for resources on speech, speech science, and speech technology.

The purpose of this paper is in part to announce the inauguration of Språkbanken Tal, in part to make mention of the services we mean to provide in the first versions of the infrastructure, and in part to present the initial formats we are putting in place for collaborations around speech research with the aim of generating dual results: research findings and publicly available resources for speech research.

Background

Språkbanken was inaugurated nearly half a century ago, in 1975, at the University of Gothenburg's department of Swedish. It has since been a nationally and internationally acknowledged research unit with a focus on language resources and language technology. Throughout the past few decades, several serious attempts have been made to implement a larger language technology infrastructure with a more sustainable funding model. Several applications to that effect have been submitted, with varying success. In 2014, Sweden received funding from the Swedish Research Council to join the European language infrastructure Clarin ERIC, which supports language technology in humanities and the social sciences, and Swe-Clarin was formed.

In 2018, the national research infrastructure Nationella Språkbanken received 105 MSEK from the Swedish Research Council, with a matching amount provided by the participating partners. Swe-Clarin is administered by Nationella Språkbanken from 2019, when its initial funding runs out.

First steps

A research infrastructure for speech is a high cost, high yield endeavour. Its design, implementation and maintenance is quite costly, and time is required for the investment to pay off. Spending the initial cost only to be forced to abandon the project would be wasteful, and our first objective is to ensure longevity. We clearly hope that Språkbanken Tal will have proven its usefulness at the end of the first funding period (2024), and that this will result in renewed funding within the programme. It would be presumptuous to take that as a given, however, and instead, we will work along several dimensions to ensure longevity early in the implementation of the infrastructure.

The first of these is to anchor the infrastructure formally on several levels. Språkbanken Tal is shortlisted as one out of nine newly established centres in the long-term initiative KTH Research

Infrastructures (KTH-RI). In 2017, KTH was the first Swedish Clarin member site to become a certified European Clarin Knowledge Centre, with speech as its domain, under the label Clarin-Speech. Finally, KTH and Språkbanken Tal make a continuous effort to complement the research infrastructure with infrastructure catering to speech industry needs as well. This is crucial for a variety of reasons. Examples include the role of speech technology in accessibility, as promoted by the Swedish Post and Telecom Authority (PTS), with whom we have undertaken resource inventories and strategic proposals (Jens Edlund, 2016, 2017), and human-centred behaviours in new technology, as seen in the strategic innovation agenda (Gustafson et al., 2014) we have developed in collaboration with VINNOVA. This parallel track of Språkbanken Tal has yet to be substantiated by funding. When that happens, we foresee significant synergy effects, as research and development often go hand in hand in speech research.

In addition to redundancy in terms of formal foundation, we work along another dimension to consolidate Språkbanken Tal at an early stage: medium to long term research collaborations with an infrastructure element written into the project plan. KTH Speech, Music and Hearing is Sweden's largest (and oldest) academic research organization for speech related research, with a project portfolio holding dozens of nationally or internationally funded, speech based research project at any given time. We hope that this position will allow us to establish several research projects together with institutions which can benefit from Språkbanken Tal to create meaningful, real-world use cases that are advantageous both to our partners and to us. In these projects, we focus on processes and methods, rather than speech data. There are a host of reasons for this decision, one of them being uncertainty surrounding the status of speech data.

Our hope is that a string of successful research projects that both make use of and add to Språkbanken Tal will not only help ensure its longevity, but also boost its design, implementation and maintenance at an early stage. We believe this to be essential for another reason as well: although the funding granted to Språkbanken Tal for the seven years between 2018 and 2024 is substantial, it is roughly but 40% of the proposed budget, and far from what is needed to properly support Swedish speech technology at a competitive level. As a result, Språkbanken Tal must household and prioritize rather harshly. We must be careful to make choices that maximize the outcome while minimizing the effort to achieve critical mass for a speech technology research infrastructure to be the active and meaningful resource it has the potential to be.

Collaborate with us!

The remainder of this announcement outlines the formats for collaboration we are currently implementing. If they seem interesting to you, please do not hesitate to contact us.

Speech technology base layer

There are a handful of speech technologies that can be considered fundamental, in the sense that they are required building blocks for a wide range of other methods and applications. As speech research is, regretfully, still rather text based, the two most obvious of these are perhaps speech-to-text, or automatic speech recognition (ASR) and speech synthesis or text-to-speech (TTS). Although there are high-quality implementations of Swedish ASR and TTS, these are exclusively commercial, and generally only available as foreign cloud services, which raises a plethora of legal concerns.

Customized ASR and TTS.

For this reason, Språkbanken Tal is looking for collaborations where a partner has a need for ASR or TTS in a specific area, and is willing to share the initial responsibilities of collecting data as well as sharing the usage data continuously. In exchange, Språkbanken Tal will adapt the ASR to increase the quality in the desired area. These adaptations are normally based on the inclusion of situation specific data into a retraining procedure.

Note that the training data to be delivered to Språkbanken Tal need not be sensitive, but can be pre-processed in various ways to make it anonymized and void of semantic coherence. A project of this type would typically require external funding in the form of a research project.

As an example, we have submitted a proposal with the National Library (KB) to adapt ASR for use with their media archive (KB).

Update sharing.

Many of the speech resources that are used in the speech technology base layer are time sensitive perishables in need of continuous (human)

updates. Språkbanken Tal will maintain repositories of such updated resources (e.g. vocabularies with frequency and pronunciation information, language and acoustic models). We are looking for collaborations in which end users of these resources who perform updates share these with Språkbanken Tal. We will open custom API's to make the delivery of updates in both directions seamless. This type of collaboration bears little cost after the initial implementation of APIs, and requires minor external funding.

By way of example, we have submitted proposals to share language model and vocabulary updates with the Swedish Agency for Accessible Media and with KB.

Advanced speech technology

Much of speech research lies beyond literal transcription and the unreflected reading aloud of texts. Speech technology, and speech research in general, is edging towards that which makes us human: the complexities of face-to-face interaction – complexities that seem so natural to us. Here, there are but few standardized methods, and we struggle with data acquisition, analysis and annotation, as well as with how to evaluate results. Språkbanken Tal seeks collaborations to develop and refine methods for annotation and evaluation of acquired speech data and multimodal data that capture human interaction.

Transcription.

A majority of the naturally occurring speech data accessible today is not a suitable match for current ASR, but must be transcribed manually at great effort and cost. There are, however, methods that can speed this process up. In particular, methods that combine automatic transcription with a human (human-in-the-loop) have shown great potential. One such method is so-called respeaking, where a speaker that is proficient in talking into an ASR repeats someone else's speech to acquire a transcription.

A very similar method is used for real-time subtitling and for interpretation. We are looking for collaborations with institutions who hold large collections of data that defies ASR, to collaborate in projects where respeaking is used to acquire the transcriptions. Språkbanken Tal will deliver the basic ASR, as well as adapt it as new data becomes available, meaning that the system becomes more efficient with usage. Respeaking and so-called shadow speaking can be used in many similar, but distinct, manners, all of which are of interest to us.

By way of examples, we have submitted a proposal to use respeaking for realtime interpretation together with Stockholm University, and a proposal to use respeaking for transliteration of handwriting with Uppsala University Library.

Experiment methods and crowd sourcing. There is a pressing need for new (and/or improved) methods to acquire reliable models of how people perceive speech and related phenomena. These methods are useful in a range of applications, from evaluation of speech technology applications, through annotation of data, to perception experiments. We are specifically looking for methods in which the participants should not be trained specialists, but everyday language users. This facilitates crowd sourcing as a method of accessing informants and affords a focus on ecological validity. In this line of research, we value ecologic validity higher than experimental control (but would ideally maximize both), so we look for perception tests that are similar to some real situation. Examples of methods we look to develop and validate for speech purposes are

- audience response systems (J Edlund & Moubayed, 2013), the screening technique used in Hollywood;
- temporally disassembled audio (Fallgren, Malisz, & Edlund, 2018) combined with massively multi-component audio environments (Jens Edlund, Gustafson, & Beskow, 2010) to present audio in clusters that are ordered differently than they originally occurred for fast browsing purposes; and
- real-time manipulation of interactions (J. Edlund & Beskow, 2009), where some aspect of e.g. a conversation is modified in realtime, and the effects of this modification are measured.

We are looking for projects in which these and other methods can be utilized, developed, and validated, and which will generate data that can be used to provide e.g. baseline models. We aim to make these methods freely available as services, and will customize the methods to fit specific projects.

As an example, we have submitted a proposal to use crowd sourcing to acquire pronunciations with Södermalms Talteknologiservice.

Data acquisition. Språkbanken Tal will not record any advanced human interaction and speech materials during its build-up phase. We will, however, participate in projects where others need to record data, and aim to adapt our resources to fit specific projects. Data acquisition projects are resource hogs, and any larger effort would require external funding. However, we will gladly offer advice or help with smaller recordings. We also aim to set up a data acquisition laboratory that can be leased complete with a technician.

As an example, we are proposing to set up a laboratory space as part of the KTH research infrastructure effort.

Conclusion

We are happy to announce that Språkbanken Tal will be built up over the next few years, and hope that it will grow into a useful resource for researchers from all fields with an interest in speech. Naturally, we hope for particularly fruitful collaborations in the phonetics area.

You can find the Språkbanken Tal portal at http://sprakbankental.speech.kth.se/

Acknowledgments

Nationella Språkbanken is funded by Swedish Research Council's programme for national research infrastructures (VR2017-00626) and by the participating parties.

References

Edlund, J. (2016). *Nyttjande av offentligt tillgängliga svenska talresurser i en nationell talresursbank*. Stockholm.

Edlund, J. (2017). *Skapandet av grunden för en svensk talbank*. Stockholm.

Edlund, J., & Beskow, J. (2009). MushyPeek: A Framework for Online Investigation of Audiovisual Dialogue Phenomena. *Language and Speech*, *52*(2–3), 351–367. http://doi.org/10.1177/0023830909103179

Edlund, J., Gustafson, J., & Beskow, J. (2010). Cocktail--a demonstration of massively multi-component audio environments for illustration and analysis. In *SLTC 2010*.

Edlund, J., & Moubayed, S. Al. (2013). Temporal precision and reliability of audience response system based annotation. *Proc. of Multimodal*.

Fallgren, P., Malisz, Z., & Edlund, J. (2018). Bringing order to chaos: a non-sequential approach for browsing large sets of found audio data. In *Proc. of the 12th International Conference on Language Resources (LREC2018)*. Miyazaki, Japan.

Gustafson, J., Edlund, J., Beskow, J., Hedelind, M., Kragic, D., Ljunggren, P., … Östlund, B. (2014). *Social robotics - a strategic innovation agenda*. Stockholm.

Productions of stop consonants, language perception and levels of agreement in interactions between Swedish adolescents

Julia Forsberg
Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg

Abstract

The present paper presents a study concerning productions of the stop consonants in the word OK in Swedish and English by Swedish adolescents, produced in map-task interaction. Speaker role in the interaction, language spoken and speaker intention will be correlated with both productions of /k/ and listener judgements on language, meaning and the pragmatic function of the utterances.

Introduction

Productions in Swedish and English by Swedish adolescents will be investigated, more specifically their utterances of the word OK in peer-paired map-tasks. These productions will be considered together with perception by listeners judging the meaning of the utterance and attempting to identify the language spoken.

Many studies concerning stop productions have traditionally focused on VOT differences and aspiration (e.g. Stölten, Abrahamsson & Hyltenstam 2015, Schmidt & Flege 1996; Kkese & Petinou 2017), and productions in Swedish have been described by Stölten, Abrahamsson & Hyltenstam 2015, Helgason & Ringen 2008 among others.

Kuronen & Zetterholm (2017:154) found that the largest negative impact on whether speakers were judged as nativelike was plosive 'errors' (unaspirated /p t k/). In Forsberg & Abelin (forthcoming), the listeners further seem to tolerate prosodic discrepancies between L1 and L2 productions more than segmental (cf. Kuronen & Zetterholm 2017:155).

Research questions on production

- In what ways are the productions of /k/ in *OK* associated with speakers' roles as information givers or information receivers?
- In what ways are the productions of /k/ in *OK* associated with the language spoken?

Research question connecting production and interpretations

- How do listener assessments of speaker intentions correspond with productions of /k/?
- How do listener assessments of language spoken correspond with productions of /k/?

Materials and method

The speech materials come from the corpus SSG which is made up of interview and map-task recordings of peer-dyads, from 111 16-19 year olds in Stockholm and Gothenburg. In each city, informants from two schools were recorded: one from the city centre, and one from a suburb. The material used for this study (so far) consists of utterances of the discourse particle OK by 6 out of the 13 female speakers from the inner city Gothenburg school. The utterances are taken from the map-task recordings, and are produced in conversation in Swedish or English. Further, each utterance is produced in one of two speaker roles: when the informant acts as information giver, or as information receiver. There are thus 4 possible conditions, each labelled as follows: SG (Swedish Giver); SR (Swedish Receiver); EG (Swedish Receiver); ER (English Receiver). Each dyad completed three map-tasks, once in each speaker role for Swedish, and either as receivers or givers for English. The only exceptions are informants D08 and D09 who due to scheduling clashes only completed one map-task. The informants were in the recordings of the map-tasks either paired with another female informant, or with a male peer (no analyses of the male speakers are included in this paper).

The productions of /k/ were transcribed using Praat (2017), and an auditory-acoustic approach. Language spoken and the speaker role were annotated. Where a plosive production was identified, the VOT was measured and the number of bursts were annotated.

The listener data takes the form of an online perception experiment where 180 listeners who reported that Swedish is their strongest language took part. The experiment was designed using Google forms, and distributed mainly through social media and e-mail lists. Listeners were asked to listen to 24 SR utterances of *OK*, and determine whether it was produced in Swedish or in English. Further, as reported in Forsberg and Abelin (forthcoming), they were asked to assess the level of agreement that was expressed through the utterance: "I agree with you" (*agree*, category 1), "I am listening, keep talking" (*go on*, category 2), "Hold on, let me think" (*hold on*, category 3) and "I am surprised by what you are saying" (*doubt*, category 4). Inferred meanings of the utterances are reflected in these four categories, based on analysis of the utterance in their interactional context, by two phoneticians. Categorisation was done individually by the two phoneticians, and there was a high level of agreement. 8 instances used in the survey were categorised as *agree* (3 English, 5 Swedish), 10 as *go on* (5 English, 5 Swedish), 2 as *hold on* (2 English) and 3 as *doubt* (1 English, 2 Swedish) by the authors.

The utterances were then correlated with language and intention, and comparisons were made with the listener judgements.

Results and analysis

Initial results will be discussed herein, with a view to extend the study to more speakers and further analysis in future papers.

Production

To date, 93 English tokens and 103 Swedish tokens have been analysed. Of the English tokens, 70 (75.3%) are produced as plosives, and 23 (24.7%) as fricatives. Of the plosive productions, 2 are voiceless palatals, 1 is a voiced velar, and the rest are voiceless velar plosives. In terms of the fricative productions in English, there are similarities to the distribution of plosives: 1 is a voiceless palatal fricative, 3 are voiced velar fricatives, and the remaining 19 are voiceless velar fricatives.

For the Swedish tokens, the division is similar, with one exception: 1 token (just under 1%) is produced as an approximant ([j]); 31 (30.1%) are fricative productions and 71 (68.9%) are plosives. Out of the plosives, 3 are voiced velars, 1 is a voiceless palatal, 1 is a voiceless uvular, and the remaining 66 are voiceless velars. The distribution of fricatives in Swedish is, again, reminiscient of both the plosives and the distribution in English: 1 is produced as a voiced palatal fricative, 2 as voiced uvular fricatives, 6 as voiceless palatal fricatives, 12 as voiced velar fricatives and the remaining 10 as voiceless velar fricatives.

Considering individual informants, it appears that the production patterns vary between speakers. Some seem to simply use more fricative productions; other more plosives. This is an area which needs further study and statistical analysis.

Listener experiments

Listeners identify the language spoken as Swedish for the majority of the utterances, both when the OK was uttered in Swedish and when it was uttered in English. There are different ways of interpreting this: either listeners who's strongest language is Swedish are simply better at correctly identifying Swedish as Swedish than English as English; or there is an expectation among listeners that the utterances will be in Swedish. As yet, no patterning of the productions of the stop consonants have been found indicating perceptual cues used by listeners in order to identify the language spoken.

Future work

Production

Not enough tokens have been analysed in order to draw conclusions on the effect speaker role (information giver/receiver) has on the production of the stop consonant. The full set of female speakers from the school will be investigated before a comparison between speaker roles can be made.

Further, the initiative of the speaker will be investigated. That is, not only the role given to the speaker in the task, but the function of the speech act in which the OK utterance occurs. The position of the utterance in the speech act will also be considered.

The productions of the stop consonants will also be further compared to the speaker intention,

and voice onset time and aspiration will be investigated in more detail.

Further investigation into the productions by the adolescent speakers depending on language will be made.

Listener experiments

The speaker intention in conjunction with the productions of /k/ will be considered in relation to listener judgements of language used.

Concluding remarks

This work-in-progress aims at connecting three areas of research: English as an L2 in Sweden; acoustic phonetic and perception analysis; and the pragmatic function of different utterances. While the focus lies on one discourse particle only, it is one that is used in interaction by adolescents (and not only adolescents) in both languages, and its various forms and functions can tell us a great deal about both interaction, phonetics and second language learning and use.

Acknowledgements

Thanks to my supervisor, Åsa Abelin, for your patience and support, to Johan Gross and Kristina Lundholm Fors for advice and valuable discussions, to Jonas Håkansson for testing the online questionnaire, and to all informants who have agreed to partake in recordings and listener experiments.

References

Boersma P & Weenink D (2017). Praat: doing phonetics by computer [Computer program]. Version 6.0.26, retrieved 2 March 2017 from http://www.praat.org/

Forsberg J & Abelin Å (forthcoming). Intonation and levels of agreement in interactions between Swedish adolescents. *Proceedings of Speech Prosody 9, 13-16 June 2018, Poznan, Poland.*

Helgason P & Ringen C (2008). Voicing and aspiration in Swedish stops. *Journal of Phonetics*, 36:607-628.

Kkese E & Petinou K (2017). Factors affecting the perception of plosives in second language English by Cypriot- Greek listeners. In E. Babatsouli (ed.), *Proceedings of the International Symposium on Monolingual and Bilingual Speech 2017*, 162-167.

Kuronen M & Zetterholm E (2017). Olika fonetiska drags relative betydelse för upplevd införddlikhet i svenska. *Nordand*, 12(1):134-156.

Schmidt AM & Flege JE (1996). Speaking Rate Effects on Stops Produced by Spanish and English Monolinguals and Spanish/English Bilinguals. *Phonetica*, 53:162-179.

Stölten KN, Abrahamsson N & Hyltenstam K (2015). Effects of age and speaking rate on voice onset time: The production of voiceless stops by near-native L2 speakers. *Studies in Second Language Acquisition*, 37:71-100.

EMA-based head movements and phrasing: a preliminary study

Johan Frid[1], Malin Svensson Lundmark[2], Gilbert Ambrazaitis[3], Susanne Schötz[4] and David House[5]
[1]Lund University Humanities Lab, Lund University [2]Centre for Languages and Literature, Lund University, [3]Department of Swedish, Linnæus University, [4]Logopedics, Phoniatrics and Audiology, Clinical Sciences, Lund University, [5]Department of Speech, Music and Hearing, KTH, Stockholm

Abstract

In this paper we describe present work on multimodal prosody by means of simultaneous recordings of articulation and head movements. Earlier work has explored patterning, usage and machine-learning based detection of focal pitch accents, head beats and eyebrow beats through audiovisual recordings. Kinematic data obtained through articulography allows for more comparable and accurate measurements, as well as three-dimensional data. Therefore, our current approach involves examining speech and body movements concurrently, using electromagnetic articulography (EMA). We have recorded large amounts of this kind of data previously, but for other purposes. In this paper, we present results from a preliminary study on the interplay between head movements and phrasing and find tendencies for upward movements occuring before and downward movements occuring after prosodic boundaries.

Introduction

This study is part of a project investigating levels of multimodal prosodic prominence, as resulting from an interplay of verbal prosody (pitch accents) and visual prosody (head and eyebrow beats). Facial beat gestures align with pitch accents in speech, functioning as visual prominence markers. However, it is not yet well understood whether and how gestures and pitch accents might be combined to create different types of multimodal prominence, and how specifically visual prominence cues are used in spoken communication.

In earlier work, Ambrazaitis & House (2017) explored the patterning and usage of focal pitch accents, head beats and eyebrow beats. The material consisted of Swedish television news broadcasts and comprised audiovisual recordings of five news readers (two female, three male). They found that head beats occur more frequently in the second than in the first part of a news reading, and also that the distribution of head beats might to some degree be governed by information structure, as the text-initial clause often defines a common ground or presents the theme of the news story. The choice between focal accent, head beat and a combination of them is subject to variation which might represent a degree of freedom for the speaker to use the markers expressively.

Based on the same, but extended data, Frid et al. (2017) developed a system for detection of speech-related head movements. The corpus was manually labelled for head movement, applying a simplistic annotation scheme consisting of a binary decision about absence/presence of a movement in relation to a word. They then used a video-based face detection procedure to extract the head positions and movements over time, and based on this they calculated velocity and acceleration features. Then a machine learning system was trained to predict absence or presence of head movement. The system achieved an F1 score of 0.69 (precision = 0.72, recall = 0.66) in 10-fold cross validation. Furthermore, the area under the ROC curve was 0.77, indicating that the system may be helpful for head movement labelling.

Kinematic vs audiovisual data

One difficulty in tackling the relationship between speech and the body gestures is that it requires simultaneously recorded kinematic and acoustic measurements. Previous studies have used audiovisual data to study this link, but with such data, it is not possible to compare synchronization of gestures directly. Kinematic data allow for more comparable and accurate measurements. Therefore, our current approach involves examining speech and body

movements concurrently, using electromagnetic articulography (EMA). This method allows for simultaneous recording of audio + 3D movements of the articulators: tongue, lips, and jaw, but markers can also be placed on the head. Head movements are typically used to normalise, but they can also be used as raw data and thereby give us the co-occurrent position of the head. Compared to video this gives us 3D coordinates instead of 2D, and has better temporal resolution (video normally has a much lower frame rate) and better audio-video sync. A disadvantage is that it must be recorded on-line; it cannot be obtained by post-processing. In this study we also employ it as an example of data reuse (Pasquetto et al. 2017): we have a large amount of data which was recorded in other projects for other purposes, but we are able to use it here to study co-occurrent properties of speech and head movements.

Data

The data was recorded as part of the VOKART project (Schötz et al. 2013). 27 native speakers (nine representing each dialect) of the Stockholm (3 female, 6 male, age: 21–63), Gothenburg (5 female, 4 male, age: 20–47), and Malmö (4 female, 5 male, age: 23–62) variants of Swedish were recorded by means of EMA using an AG500 (Carstens Medizinelektronik) with a sampling frequency of 200 Hz. Ten sensors were attached to the lips, jaw and tongue, along with two reference sensors on the nose ridge and behind the ear to correct for head movements, using Cyano Veneer Fast dental glue. Audio was recorded using a Sony ECM-T6 electret condenser microphone.

The speech material consisted of 15–20 repetitions by each speaker of target words in carrier sentences of the type “Det va inte hV1t utan hV2t ja sa” (It was not hV1t, but hV2t I said), where V1 and V2 were different vowels. The target words containing the vowels were stressed and produced with contrastive focus. The sentences were displayed in random order on a computer screen, and the speakers were instructed to read each sentence in their own dialect at a comfortable speech rate. In order to familiarise the speakers with the sensors and the experimental setup the actual test sentences were preceded by two phonetically rich and challenging sentences, which the speakers were asked to repeat three times each. The two sentences were:

1) *Mobiltelefonen är nittiotalets stora fluga, både bland företagare och privatpersoner.* (The mobile phone is the big hit of the nineties, both among business people and private persons.)

2) *Flyget, tåget och bilbranschen tävlar om lönsamhet och folkets gunst.* (Airlines, train companies and the automobile industry are competing for profitability and people's appreciation.)

In addition, the speakers were also asked to describe a painting displayed on the computer screen, resulting in about half a minute of spontaneous speech, with several focused words and phrase-boundaries. A contour of the palate was obtained by the speakers moving their tongue tips several times back and forth along the midline of their palate.

Analysis: head movements and phrase boundaries

We analyze the data by looking at sentence-level patterns of head movements and comparing them word by word. Sentence 1 above consists of two phrases, with an intonational boundary between the words *fluga* and *både*. Sentence 2 is essentially one phrase, but starts with a list that may cause boundary signalling. We have examined parts of the material for possible head movement reflections of those boundaries. In order to get an annotation of the word boundaries of the sentences, we use the alignment method provided by the Praat program (Boersma & Weenink 2018), which speeds up the process but still requires manual post-checking. At the time of writing, we have analysed and checked the data from twenty-four of the speakers (originally three repetitions each). Utterances that 1) contained misreadings, 2) contained missing parts (because the recording stopped before the reader finished the sentence), and 3) have not yet had their alignment checked, were discarded. In the present study we have 61 examples of Sentence 1 and 44 examples of Sentence 2.

First we measured the velocity of the angle in the sagittal plane between 1) an imaginary line between the two reference sensors (behind the ear and on the nose ridge) and 2) a line running along the transverse plane (parallel to the ground). This effectively measures the head's movement as it is tilted along this plane. We then calculated the average angular velocity per

word for each sentence. Figures 1 and 2 show summaries of the data in the form of boxplots. For Sentence 1 (in Figure 1), we note that the boundary-preceding word *fluga* has a positive median, whereas the the following word, *både*, has a negative median. A similar, but less prominent, pattern can be observed in Sentence 2 (Figure 2), where the first word *Flyget* has a positive median, whereas the following word *tåget*, has a negative median.

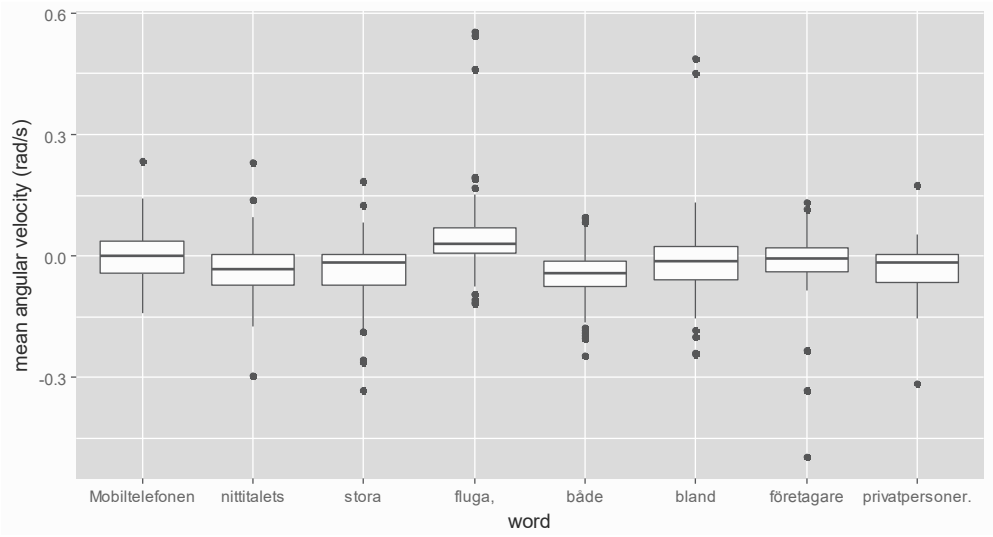

Figure 1. Boxplots of mean angular velocity per word in sentence 1, n=61. Black horizontal lines are medians, hinges correspond to the first and third quartiles, black dots are outliers.

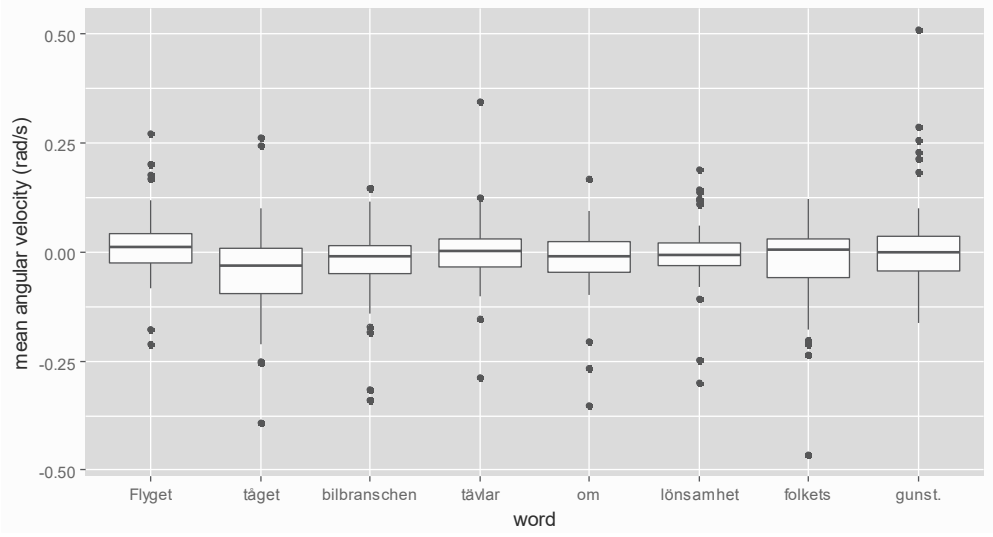

Figure 2. Boxplots of mean angular velocity per word in sentence 2, n=44. Black horizontal lines are medians, hinges correspond to the first and third quartiles, black dots are outliers.

A paired-samples t-test was conducted to compare the average angular velocity for the words. In Sentence 1, there was a significant difference in the scores for *fluga* (M=0.056,

SD=0.126) and *både* (M=-0.046, SD=0.076) conditions; $t(60)=4.706$, $p < 0.05$, and in Sentence 2, there was a significant difference in the scores for *Flyget* (M=0.01, SD=0.09) and *tåget* (M=-0.04, SD=0.113) conditions; $t(43)=2.0462$, $p < 0.05$.

Discussion/Conclusions

Using EMA recordings to analyze head movement by comparing the kinematic patterns of the sensors with the audio signal is a promising method to provide us with information on the synchronization of head movements with for example prosodic signals for prominence such as F0 excursions and syllable lengthening.

The results presented here show that there is a tendency for participants to tilt the head more upwards than downwards during the boundary-preceding words. The words which succeed the boundary, conversely show an opposite tendency indicating more downward movement.

Previously motion capture data has been used to investigate temporal coordination between head movement and the audio signal (Alexanderson et al. 2013) and between head movement and EMA articulation data (Krivokapić et al. 2017; Esteve-Gilbert et al. 2018). EMA methodology has also been used to analyze head movements alone, but the current data will enable us to investigate the temporal coordination of head movements, tongue and lip movements, and the audio signal in the same system. We plan to use this methodology to investigate the role of head movement, articulation and prosody in signaling prominence in the context of the newly initiated PROGEST project.

Acknowledgements

This work was supported by grants from the Swedish Research Council: Swe-Clarin (VR 2013-2003) and Progest (VR 2017-02140).

References

Alexanderson, S, House, D, & Beskow, J. (2013). Aspects of co-occurring syllables and head nods in spontane-ous dialogue. In *Proc. of 12th International Conference on Auditory-Visual Speech Processing* (AVSP2013). Annecy, France.

Ambrazaitis, G., & House, D. (2017). Multimodal prominences: Exploring the patterning and usage of focal pitch accents, head beats and eyebrow beats in Swedish television news readings, *Speech Communication*, 95, pp. 100-113, https://doi.org/10.1016/j.specom.2017.08.008

Boersma, Paul & Weenink, David (2018). Praat: doing phonetics by computer [Computer program]. Version 6.0.39, retrieved 3 April 2018 from http://www.praat.org/

Esteve-Gibert, N., Loevenbruck, H., Dohen, M. & D'Imperio, M. (2018) Head movements highlight important information in speech: an EMA study with French speakers. DOI10.13140/RG.2.2.21796.78727 *Conference: XIV AISV Conference - Speech in Natural Context*, 25-27 January 2018, Bozen-Bolzano

Frid, J., Ambrazaitis, G., Svensson-Lundmark, M. & House D. (2017). Towards classification of head move-ments in audiovisual recordings of read news, *Proceedings of the 4th European and 7th Nordic Symposium on Multimodal Communication* (MMSYM 2016), Copenhagen, 29-30 September 2016, Volume, Issue 141, 2017-09-21, Pages 4-9, ISSN 1650-3740

Krivokapić, J., Tiede, M.K. & Tyrone, M. E. (2017). A Kinematic Study of Prosodic Structure in Articulatory and Manual Gestures: Results from a Novel Method of Data Collection. *Lab Phonol.* 2017; 8(1): 3. Published online 2017 Mar 13. doi: 10.5334/labphon.75

Pasquetto, I.V., Randles, B.M. & Borgman, C.L., (2017). On the Reuse of Scientific Data. *Data Science Journal.* 16, p.8. DOI: http://doi.org/10.5334/dsj-2017-008

Schötz, S., Frid, J., Gustafsson, L., & Löfqvist, A. (2013). Functional Data Analysis of Tongue Articulation in Palatal Vowels: Gothenburg and Malmöhus Swedish /iː, yː , ʉː/. *Proceedings of Interspeech 2013*. Lyon.

Exploring Voice Quality Changes in Words with Stød

Gert Foget Hansen
Department of Nordic Studies and Linguistics, University of Copenhagen

Abstract

Stød is a prosodic feature occurring in Danish. The most conspicuous acoustic trait of stød in its prototypical form is a short stretch of irregular vocal fold vibrations, i.e. creak. However, creak is neither necessary nor sufficient to characterize stød: The occurrence of creak is not limited to syllables with stød and distinct and clear realizations of stød need not exhibit creak.

To account for the inconsistent occurrence of irregular vocal fold vibrations in stød it is hypothesized that stød could be explained as a relative and dynamic voice quality movement in the form of a brief change from less to more compressed voice, potentially but not necessarily involving creaky voice.

To test the hypothesis changes in voice quality are traced over the cause of comparable syllables with and without stød using a set of voice quality related acoustic measures.

The results demonstrate that the timing of the peak level of compression need not coincide with the occurrence of irregular vibrations. As a consequence of these findings the proposed stød hypothesis is rejected. Moreover, the results challenge the underlying models of voice quality, as results do not conform to generally accepted assumptions about the relation between creaky voice and compression.

Introduction

The Danish stød is a syllable prosody, that may occur in certain syllables, namely stressed syllables with a long vowel or with a short vowel followed by a sonorant consonant. At the surface level stød often distinguishes meaning, even though the distribution of stød is fairly predictable from albeit rather complex morpho-syntactic patterns, see for instance Basbøll (1985) and Basbøll (1998).

Phonetically the stød has been described as "a kind of creaky voice" (Grønnum, 1998, p. 103) though the stød is notorious for its elusive phonetic characteristics, see for instance Petersen (1973), Fischer-Jørgensen (1989), and Grønnum & Basbøll (2007).

The problem with describing stød as (a kind of) creaky voice is that irregular vocal fold vibrations occurs rather often in stød, but with quite a range of variation: Some speakers on some occasions rarely produce stød without irregular vibrations, while some speakers on some occasions, rarely produce stød with irregular vibrations. Notably, stød need not exhibit irregular vibrations to be perceived as distinct and clear realizations, and the occurrence of irregular vibrations is not limited to syllables with stød. The occurrence of irregular vibrations is thus neither necessary nor sufficient to characterize stød.

That being said, irregular vocal fold vibrations is seen as emblematic for a prototypical stød. It has been suggested that the presence or absence of irregular vocal fold vibrations in stød could be a matter of more or less distinct realization, see Fischer-Jørgensen (1989) and Grønnum & Basbøll (2007). However, numerous counterexamples to such an interpretation have been given, cf. Petersen (1973), Thorsen (1974), and Grønnum & Basbøll (2007), and it is evident that stød which does not exhibit irregular vocal fold vibrations may sound as strong and may be perceived as readily as stød exhibiting irregular vocal fold vibrations.

A drop in intensity appears to be the most consistent acoustic trait of stød. The intensity drop is most likely the result of a glottal constriction, cf. Fischer-Jørgensen (1989). It must be emphasized that a true glottal closure rarely occurs in connection with stød.

Many sources report seeing a local drop in f_0. There is reason to suspect that most of these reports are not due to real changes in f_0 but are artifacts of the (various) pitch tracking methods

used and due to the difficulties inherent in determining f_0 when the vocal fold vibrations become less regular. Based on narrow band spectrograms (which is a more robust method in these cases) actual f_0 drops can sometimes be observed in connection with stød, but it seems to be the exception rather than the norm.

Hypothesis

In the following a model for the stød which is able to handle the inconsistent presence of creak without linking it to more or less distinct realization of the stød will be presented. The vantage point for the model is the ranging of voice qualities according to glottal stricture as presented in Ladefoged (1971) as well as Gordon & Ladefoged (2001); see table 1.

Table 1. Voice qualities ranged according to glottal stricture, based on Ladefoged (1971).

Glottal opening	Phonetic term		Compression
Closed	Glottal stop		
·	Creak		
·	Creaky voice		
·	Pressed voice	↑	Hyperfunctional ~ lower OQ
·	**Modal voice**	·	**Ideal/optimal**
·	Breathy voice	↓	Hypofunctional ~ higher OQ
·	Whisper		
Open	Breathing		

The idea is that stød is expressed through a relatively brief change in voice quality towards a more compressed, possibly creaked voice quality and subsequently back towards less compressed voice. The concept is illustrated in figure 1. This way stød is seen as a relative and dynamic voice quality gesture. A well formed (distinct) stød is presumably expressed by a suitably large change in voice quality over a suitably short span of time. Depending on how high up in the range of voice qualities the starting point is, a distinct realization of a stød may or may not include a stretch of creak.

Conceptually this is analogous to how tonal patterns may be described. Conveying a tonal pattern usually isn't about reaching a certain tone, but rather about producing a particular tonal pattern pattern with a suitable range and timing it precisely to the relevant phonetic units.

In short: Where tonal patterns such as for instance the danish stress group pattern or Swedish word tones play out in the tonal domain, the stød is proposed to play out in a 'voice quality' or timbre domain.

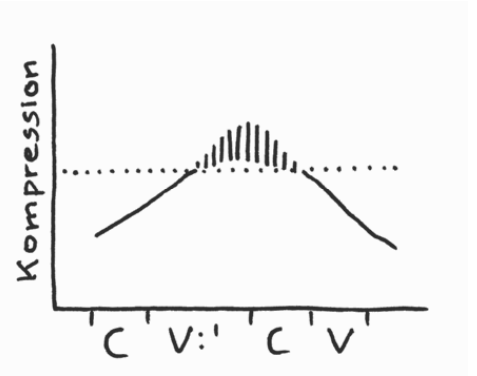

Figure 1. Illustration of the hypothesized compression pattern. V:' indicates long vowel with stød. The solid rising – falling line illustrates the increasing – decreasing compression, whereas the passage with vertical lines illustrates irregular vocal fold vibrations. The dotted horizontal line illustrates the idea of a boundary between regular and irregular vocal fold vibrations depending on compression.

Operationalization

The stated hypothesis does not lend itself directly to empirical testing but, following the concept laid out in the model and drawing on Ladefogeds hierarchy, some well established assumptions on the relation between irregular vocal fold vibrations and compression, and knowledge about the cause of the intensity dip seen in connection with stød, cf. Fischer-Jørgensen (1989), one can state four sub hypotheses that can be tested using acoustic phonetic methods.

The sub hypotheses are stated with reference to three (types of) quantifiable acoustic measures:

A. Measures related to spectral balance which reflect the vocal compression.
B. Periodicity
C. Intensity

The relation between the acoustic measures and the four sub hypotheses (stated below) are illustrated in figure 2.

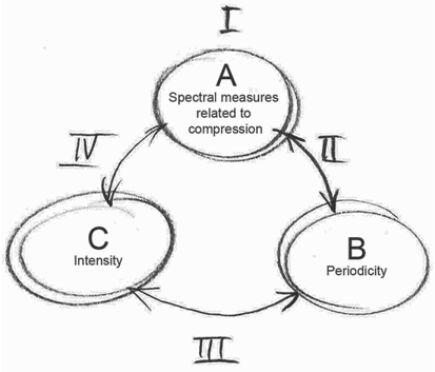

Figure 2. Illustrating the relation between the four sub hypotheses (I-IV) and the acoustic correlates for compression (H1:H2, BED and CoG), periodicity, and intensity (A-C).

I. The acoustic measures related to spectral balance – reflecting the vocal compression – are expected to follow a rising–falling trajectory. In cases where the stød is followed immediately by unvoiced material the trajectory could be truncated to show only a rising trajectory.

II. The acoustic measures related to spectral balance – reflecting the vocal compression – are expected to indicate higher compression in stretches showing irregular vibrations relative to stretches with regular vibrations; i.e a correlation between higher compression and the occurrence of irregular vibrations

III. Whenever stød is realized with irregular vocal fold vibrations the timing of the intensity minimum which is taken to signify the timing of the strongest glottal constriction is expected to occur within the stretch of irregular vibrations.

IV. The acoustic measures related to spectral balance – reflecting the vocal compression – are expected to indicate the highest level of compression near the time of the intensity minimum which is taken to signify the timing of the strongest glottal constriction.

A token must be in accord with all of these four sub hypotheses to support the main hypothesis.

Implementation

The main obstacle lies in how to gauge compression in and around passages with irregular vocal fold vibrations. The H1 to H2 ratio is a measure often used, but this measure has obvious limitations: In the case of genuinely irregular vocal fold vibrations there are no harmonics to measure (as harmonics are a property of periodic signals). Because of this, all acoustic measures used to characterize voice quality which depends on detecting the strength of particular harmonics are unfit to describe voice quality when the vibratory pattern becomes irregular.

The diplophonic vibration patterns, which quite often precede or follow genuinely irregular vibrations, pose separate difficulties for measures relying on the strength of particular harmonics, see analysis in Hansen (2015).

A more compressed voice (with a lower Open Quotient) has a less steep spectral slope, with relative more energy in the upper parts of the spectrum, corresponding with an auditory impression of a sharper timbre, whereas less compressed voice (with a higher Open Quotient) has a steeper spectral slope, showing less energy in the upper parts of the spectrum, corresponding with an auditory impression of a softer timbre. Assuming that these links between OQ and spectral slope holds whether the vocal fold vibrations are periodic or not, other ways of gauging the spectral slope of the speech signal can be used to assess the vocal compression in passages with irregular vocal fold vibrations.

Thus it was decided to use a combination of measures. H1:H2 is measured for stretches where f_0 can be detected. The H1:H2 ratio is combined with two less standard measures.

One is the measure of spectral Center of Gravity for the frequency range 1-500 Hz. Band limiting the signal was found to remove some unwanted contamination from fricative noise in neighbouring segments. CoG was found to perform better than spectral tilt. The main advantage of the CoG based measure over H1:H2 is that it does not rely on the presence or precise detection of harmonics to be calculated.

The second measure, termed BED for Band Energy Difference compares the level energy in two distinct bands ranging from 1 Hz to 1.5 times H1 and 1.5 times H1 to 500 Hz. The 1.5 x H1 is based on an interpolated f_0-tracking for those stretches where there are no harmonics or where the f_0-algorithm fails to track f_0. BED is in a sense intermediate between H1:H2 and CoG in that it does require an estimate of f_0 to be computed, but it is more lax with regards to the required precision of the f_0 estimate. See Hansen (2015) for further details.

All three measures are based on the output spectrum. As a consequence the H1:H2, CoG and BED values are affected by the interplay

between the voice source spectrum and the vocal tract resonances. Thus changes in pitch and predominantly the lowest vocal tract resonances, will affect the obtained values – more so in cases of high pitch and/or low F1. This issue is mitigated to some degree by comparing curves between minimal word pairs with and without stød. The combination of three measures also helps mitigate this issue, since H1:H2, BED and CoG are affected by the voice source/vocal tract filter interplay in somewhat different ways. See Hansen (2015) for further details.

Determining the periodicity or regularity of the vocal fold vibrations can be done in a number of ways. Here a simple distinction is made between regular and irregular (including diplophonic) vocal fold vibrations based on visual inspection of narrow band spectrograms; as the vibratory pattern becomes irregular the breaking down of the harmonic structure is visibly clear.

Intensity is measured using the intensity function in Praat.

Test

The hypothesis is tested on a material consisting of 238 tokens – 136 tokens with stød, and 102 tokens without stød – covering a range of contexts for the stød. This includes a number of vowel qualities [iɛæɑ], vowels with stød vs sonorant consonants with stød [lm], and stød in syllables with primary vs secondary stress. The material is based on sentences read aloud by one male speaker, recorded under laboratory conditions. Testing on a larger material, using more speakers was planned, but the present material proved sufficient to dismiss the presented hypothesis.

A large part of the material consists of minimal (or subminimal) pairs of words with and without stød, in order to facilitate a comparison of curves between tokens with and without stød to gauge the effects on the measured H1:H2, BED and CoG values caused by the expected voice quality differences over the base values influenced by pitch and vocal tract resonances.

For some tokens the chosen methodology turned out to be inadequate as it was not possible to accurately access whether the observed changes in the voice quality related measures was due to actual differences in the voice quality or whether they were due to differences in f_0 and/or vocal tract resonances. This was in particular the case for many tokens with the high vowel [i] and for most tokens with sonorant consonants [lm] due to very low first resonance frequencies interacting with H2. It was found that in such cases very small changes in f_0 or vocal tract resonances can have a profound effect on the H1:H2, BED and CoG values. Tokens with an [æi] diphthong, also proved difficult to analyze due to the dynamicly changing F1.

Some tokens presented patterns in good agreement with the stated hypothesis. One such example is given in figure 3. H1:H2, BED and CoG all show an upwards trend during the course of the vowel, and regarding BED and CoG continuing this trend into the irregular phase culminating near the intensity minimum.

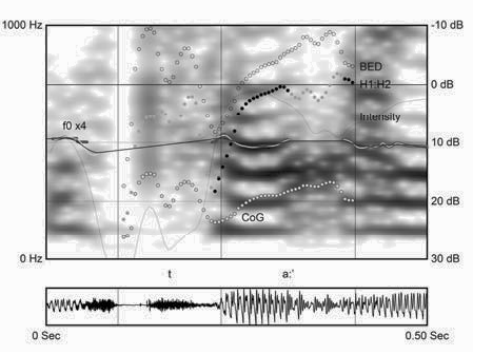

Figure 3. Curves for word with stød which is in agreement with the hypothesis. Note that the axes for the H1:H2 and BED is reversed so that more compressed voice quality plots higher in the graph, just as with CoG.

Tokens without stød shows much more level curve trajectories. One example shown in figure 4. Compare with the minimal pair sibling with stød in figure 5.

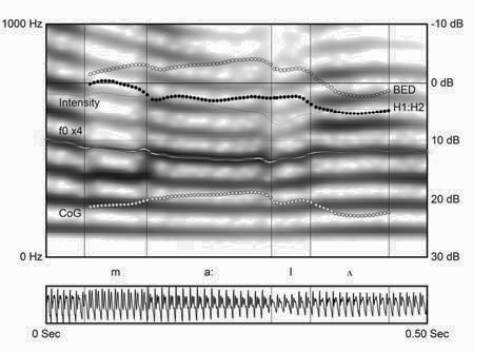

Figure 4. Curves for word without stød.

However, many tokens with stød display curve trajectories that are not in accord with the stated hypothesis. One such example is given in figure 5. H1:H2, BED and CoG all show maxima before the irregular vocal fold vibrations sets in, and thus a downwards trend leading into the stretch of irregular vocal fold vibrations.

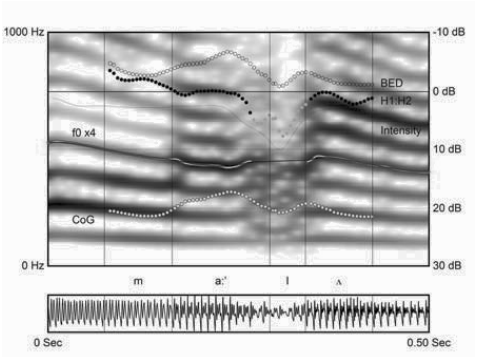

Figure 5. Currves for word with stød which does not comply with the stated hypothesis.

The voice quality measures do show a rising falling pattern, and the intensity dip does coincide with the occurrence of irregular vocal fold vibrations, but the voice quality measures indicates a maximum that falls outside the stretch of irregular vocal fold vibrations and not near the intensity minimum. Thus this sample contradicts two of the four sub hypotheses.

Results

Table 2 tallies up how the 136 tokens comply with the sub hypotheses and the main hypothesis. Sub hypothesis III is not included in the table, since all tokens were in agreement with that sub hypothesis. I.e. in all instances the intensity minimum (which is taken to signify the timing of the strongest glottal constriction) occurred within the stretch of irregular vibrations.

A number of tokens contradict one or more of the other sub hypotheses:

Regarding sub hypothesis I: Of the 136 tested tokens there are 8 cases where the acoustic measures reflecting the vocal compression can be said not to follow the expected rising–falling trajectory or, in cases where the stød is followed immediately by unvoiced material the trajectory, a rising trajectory.

Table 2. Results for the 136 stød tokens regarding the sub hypotheses (I, II & IV) and the main hypothesis.

n=136	Sub hypotheis			Main hypothesis
	I	II	IV	
Positive	72 (53 %)	26 (19 %)	37 (27 %)	22 (16 %)
Negative	8 (6 %)	31 (23 %)	42 (31 %)	52 (38 %)
Uncertain	56 (41 %)	79 (58%)	57 (42 %)	62 (46 %)

Regarding sub hypothesis II: In 31 cases the acoustic measures reflecting the vocal compression indicates that the highest level of compression occurs before the stretch of irregular vibrations.

Regarding sub hypothesis IV: In 42 cases the acoustic measures reflecting the vocal compression indicates that the highest level of compression does not coincide with the timing of the intensity minimum taken to signify the timing of the glottal constriction.

In total 52 out of the 136 stød tokens contradict the main hypothesis by contradicting one or more of the four sub hypotheses. As a consequence the proposed model is rejected.

Conclusions

The results demonstrate that the timing of the peak level of compression need neither coincide with the occurrence of irregular vibrations nor with the intensity minimum indicating the time of the strongest glottal constriction in stød. As a consequence of these findings the proposed stød hypothesis is rejected.

The fact that some tokens presents patterns in good agreement with the stated hypothesis, while other tokens presents patterns that contradict it can only mean that the changes in voice quality seen in relation to stød can unfold in more ways than one which includes irregular vocal fold vibrations.

Thus, rather than contributing to solving the puzzle of why apparently equally distinct realizations of stød can show such a bewildering variation in its acoustic presence, it uncovers yet another type of variation in the way stød may present itself.

Moreover, some of the results challenge the underlying models of voice quality, as the

results do not conform to the generally accepted assumptions about the relation between compression and the occurrence of irregular vocal fold vibrations.

References

Basbøll H (1985). Stød in modern Danish. *Folia Linguistica*, 19: 1-50.

Basbøll H (1998). Nyt om stødet i moderne rigsdansk – om samspillet mellem lydstruktur og ordgrammatik. In: Kjær & Lundgreen-Nielsen, eds, danske studier *1998*. Denmark: C. A. Reitzels Forlag, 33-86.

Fischer-Jørgensen E (1989). *A Phonetic Study of the Stød in Standard Danish*. Finland: University of Turku.

Gordon M & Ladefoged P (2001). Phonation types: a cross-linguistic overview. *Journal of Phonetics*, 29: 383-406.

Grønnum N (1998). Danish. *Journal of the International Phonetic Association*, 28: 99-105.

Grønnum N & Basbøll H (2007). Danish Stød – Phonological and Cognitive Issues. In: Solé, Beddor & Ohala, eds, *Experimental Approaches to Phonology*. United Kingdom: Oxford University Press, 192-206.

Hansen G F (2015). *Stød og stemmekvalitet*. Denmark: unpublished PhD thesis, http://static-curis.ku.dk/portal/files/136723718/Ph.d._2015_Foget_Hansen.pdf 2015_Foget_Hansen.pdf

Ladefoged P (1971). *Preliminaries to Linguistic Phonetics*. USA: The University of Chicago Press.

Petersen P R (1973). An instrumental investigation of the Danish "stød". *ARIPUC*, 7: 195-234.

Thorsen N G (1974). Acoustical and perceptual properties of the Danish stød. *ARIPUC*, 8: 207-213.

'Fax it up.' – 'Yes it does rather.' The relationship between the TRAP, STRUT, and START vowels in Aberystwyth English

Míša Hejná
Department of English, Aarhus University

Abstract

This paper focuses on the relationship between the TRAP, STRUT, and START phonemes in English spoken in Aberystwyth, mid Wales. It concludes that, in general, whilst the three phonemes do not show a complete merger, they do overlap to a great extent regarding a number of acoustic properties considered. It is predominantly F1 (for TRAP and STRUT) and vowel duration (for TRAP and START, and STRUT and START) that serve as correlates of the contrasts rather than F2 and the breathiness of the vowel. However, TRAP and STRUT do show a high degree of overlap in F1.

Unimodality tests reveal that TRAP and STRUT are associated with two independent modes regarding the distribution of F1 only in one of the ten speakers analysed. This does not apply to F2, vowel duration, and breathiness, where separate modes do not emerge in the statistical analyses for any speakers. On the other hand, a random forest analysis suggests that F1 is the most important correlate of the TRAP-STRUT contrast on the whole.

Regaring the START-TRAP and START-STRUT contrasts, unimodality tests show that START is associated with an independent mode concerning vowel duration for three of the ten speakers. A random forest analysis corroborates that vowel duration is generally the most important correlate of these two contrasts.

TRAP-STRUT merger

Crystal (2013) presents us with the following accent-related joke: "A judge arrives at his chambers having left an important document at home. 'Fax it up', his clerk suggests. 'Yes it does rather', replies the judge." Whilst this is presented as misinterpretation due to differences across accents, some researchers have noted a potential overlap between the TRAP and the STRUT vowels within a single accent. This has been suggested for RP by Wells (1982: 291-292), who comments on the increased perceptual similarity of the two phonemes in RP and who speculates that "[i]t may even be the case for some speakers that /æ/ and /ʌ/ are merged, variably at least." (1982: 292) Fabricius (2007: 311) has provided a quantitative production study that shows that TRAP and STRUT have indeed been changing regarding their absolute values in RP. However, Fabricius argues that these changes do not show configurational differences. Yet, some of the figures shown in Fabricius (2007) suggest a decrease in the difference between the phonemes on the F1 dimension.

Interestingly, RP is not the only accent for which a potential TRAP-STRUT merger has been proposed. An overlap between the two phonemes has also been reported for "some Southwestern Scots speakers" (Hall-Lew et al. 2017: 345). In addition, and crucially for the purposes of this paper, Hejná (2015: 271-3) has noted that there may indeed be a merger of the TRAP and the STRUT vowels (so that *fax* may sound like *fucks*) in English spoken in Aberystwyth, mid Wales (WE is used for Welsh English). These reports may imply a geographically relatively widespread potential merger within the UK.

Whilst Fabricius' work presents an essential instrumental study related to the subject-matter at hand, the incidental findings for Aberystwyth English, which point mainly to a merger in the F2 dimension, have not been examined in detail.

This paper therefore presents an analysis aimed to answer primarily the following question:

1. Is there a TRAP-STRUT merger in Aberystwyth English? More specifically, is there a merger in F1 and/or F2 of the TRAP-STRUT phonemes?

Although this study targets what may be a geographically limited phenomenon from a more global point of view (in comparison to, for example, High Rising Terminals – e.g. Britain 1992, Fletcher et al. 1999, Ritchart & Arvaniti 2014, Sando 2009, Shobbrook & House 2003, Shokeir 2008; GOOSE-fronting – e.g. Price 2008, Strycharczuk & Scobbie 2017, Ward 2003; and glottalisation – e.g. Cox & Palethorpe 2007: 342-343, Eddington & Channer 2010, Eddington & Taylor 2009, Foulkes & Docherty 1999, Mees & Collins 1999), any study of a potential merger enables us to explore the relationships between a number of acoustic dimensions relevant for vowel contrast beyond the traditional F1 and F2 dimensions. Importantly, in the studies of mergers, it has been pointed out that what may appear as full mergers when approached from F1 and F2 analyses may in fact constitute near mergers if the contrast is preserved via durational differences of the two vowels (see e.g. Labov & Baranowski 2006) or phonation (Di Paoli 1990). Links between phonatory settings and what is primarily thought of as oral phenomena have been shown and suggested (Brunner & Żygis 2011, Lotto et al. 1997); however, their insights remain to be extended to the field of sociolinguistics. This evidence motivates the following question:

2. Do vowel duration and phonation contribute to the phonetic implementation of the TRAP-STRUT contrast?

Answering the question will contribute to our understanding of the potential multiplicity of correlates of vowel contrasts, which have received less attention than consonants in this respect (e.g. Al-Tamimi & Khattab 2011, Toscano & McMurray 2010).

TRAP and START in WE

Whilst the START vowel has a backer quality in SSBE (Wells 1982), which makes it fairly distant from TRAP regarding F2, in WE it has been traditionally described as central (Wells 1982: 380-3), presumably due to the effects of Welsh. Additionally, the TRAP phoneme has been undergoing retraction in some British English accents (Wells 1982: 356, 380-383, 386). Considering that both TRAP and START can be central in WE, the following questions suggest themselves:

3. What are the phonetic correlates of the TRAP-START contrast?
4. What are the phonetic correlates of the STRUT-START contrast?

Methodology

This section introduces the social characteristics of the speakers, the structural properties of the material analysed, the technical aspects of the recording process, and finally the analysis of vowel formants, vowel duration, and breathiness.

Speakers

This study uses production data from 10 female participants, all of whom are native L1 Welsh speakers born and raised in Aberystwyth, mid Wales. The production data is based on their English production. At the time of the recording (2012-2013), the speakers' age ranged between 24 to 89 years. Age does not affect the results.

Data

The speakers read 'CVC and 'CVCV words, each type once in isolation (*hat* / *hut* / *heart*) and twice in a carrier sentence (*Say hat* / *hut* / *heart once*.). This was part of a larger project targeting fortis obstruents and the data analysed in this study represents only a subpart of the entire dataset. The segmental and the prosodic conditions are identical across the speakers. The second consonant is always a fortis obstruents (/p/, /t/, /k/ for all three vowels; /θ/, /s/, /ʃ/ for TRAP and STRUT; and /f/ for STRUT). In total, 1295 TRAP tokens, 745 STRUT, and 471 START tokens were included in the present analyses. For more details, see Hejná (2015: 308-312).

Procedure

The speakers were reading the target words and sentences in a randomised order. They were wearing a head-mounted AKG C520 microphone and the recording device used was H4 Zoom Handy Recorder.

Statistical analyses were conducted in R (R Core Team 2018) and RStudio, using the lme4 (Bates et al. 2015), ranger (Wright et al. 2018,

Wright & Ziegler 2015), and scatter3d (Ligger & Mächler 2003) packages.

Formant measurements

Formant measurements were extracted automatically in the midpoint of the modal interval of the vowel with the default 25ms analysis window in Praat (Boersma & Weenink 2014). The measurements included F1, F2, and F3, and were all checked manually for potential measurement errors. Different formant settings were used for the individual speakers, as appropriate. Tokens with whispered (i.e. voiceless) and exceedingly breathy vowels were excluded from the analyses (see Hejná 2015: 136 for more details).

Normalised values of the measurements were obtained via the NORM suite (Thomas & Kendall 2007-2014). Nearey formant intrinsic normalisation was used (Nearey 1977).

Durational measurements

The onset of the vowel and its offset were identified on the basis of the periodicity onset and offset through the inspection of the waveform. Vowel segmentation was conducted manually. Normalised vowel duration (quantified as a percentage of the overall word duration) provided the same results as raw vowel duration.

Brethiness measurements

The phonation characteristic of interest in the present study is that of breathiness, since the speakers produce glottalisation fairly infrequently in the data (excepting one of the speakers included in this study). CPP was therefore used to quantify the amount of breathiness in the vowels, as it has been identified as the best correlate of perceived breathiness (Hillenbrand, Cleveland & Erickson 1994: 776, Fraile & Godino-Llorente 2014). CPP measurements were extracted in VoiceSauce (Shue 2010; Shue, Keating, Vicenik & Yu 2011) in Matlab (2016).

Results

TRAP-STRUT formants

Pooling the data across all ten speakers, we find that the visual inspection of the distribution of F1 is bimodal (Fig. 1) and that of F2 is unimodal (Fig. 2). Hartigans' Dip Test (Hartigan & Hartigan 1985, Maechler 2015) nevertheless does not confirm the presence of a bimodal distribution in either case (F1: D = 0.005549, p-value = 0.9824; F2: D = 0.0061765, p-value = 0.9295).

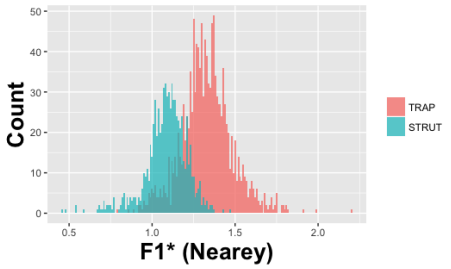

Figure 1. Distribution of F1 (normalized) for TRAP and STRUT.

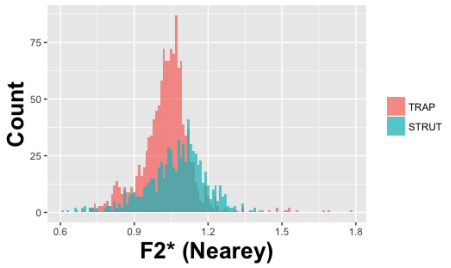

Figure 2. Distribution of F2 (normalized) for TRAP and STRUT.

Narrowing the analyses to the specific individuals, only one shows a bimodal distribution in F1 (ABE24; D = 0.04, $p < 0.01$), whilst the rest exhibit a unimodal pattern (D = 0.01-0.04, p = 0.11-0.996). Regarding F2, none of the speakers shows a bimodal distribution (D = 0.01-0.03, p = 0.45-0.995).

TRAP-STRUT duration

Pooling the data across all ten speakers again, we find that the distribution of vowel duration is unimodal (Fig. 3), and Hartigans' Dip Test confirms this unimodality (D = 0.01, p-value = 0.9922). A closer inspection of the individual speakers does not reveal any bimodality either (D = 0.02-0.03, p = 0.39-0.95).

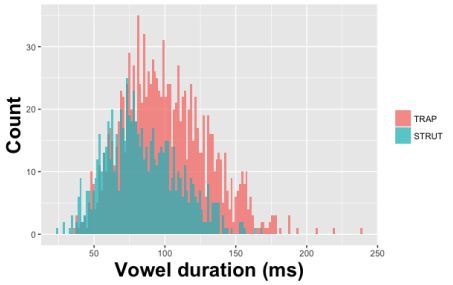

Figure 3. Distribution of vowel duration for TRAP and STRUT.

Place of START: formants

The distribution of F1 of the START vowel is found at the intersection of the STRUT and the TRAP phonemes (Fig. 4) and does not present a statistically confirmed distinct mode (D = 0.005, p = 0.98). Visually, START is associated with lower F2 than the other two vowel phonemes (Fig. 5), although this is not confirmed by the statistical analysis (D = 0.005, p = 0.97).

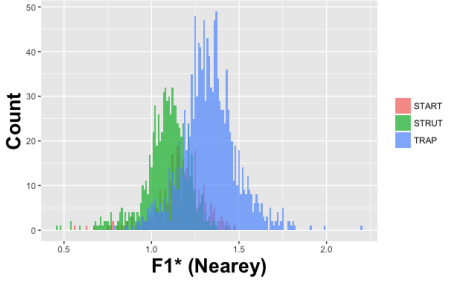

Figure 4. Distribution of F1 (normalized) for START with respect to TRAP and STRUT.

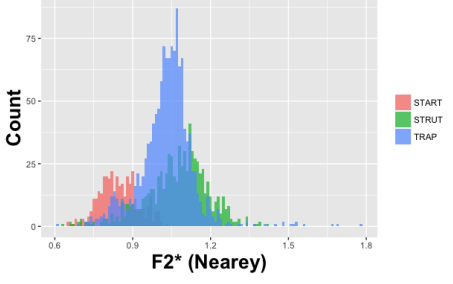

Figure 5. Distribution of F2 (normalized) for START with respect to TRAP and STRUT.

Individual variation is found only with respect to F1, for which ABE24 shows bimodality again (D = 0.03, p < 0.05). The remaining speakers suggest unimodal distributions (D = 0.01-0.03, p = 0.17-0.99). Considering F2, none of the speakers' distributions are bi- or multimodal (D = 0.02-0.03, p = 0.1-0.98).

Place of START: duration

Pooling all the speakers together, Fig. 6 shows that the START vowel is associated with fairly different durations than either STRUT or TRAP, although the Hartigans' Test rather surprisingly does not confirm this trend (D = 0.004, p = 0.99).

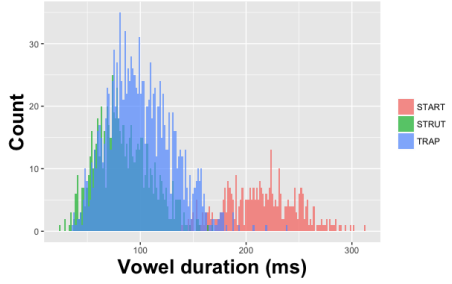

Figure 6. Distribution of vowel duration (ms) for START with respect to TRAP and STRUT.

Two individuals reveal a significant bimodal distribution (ABE31: D = 0.04, p < 0.05; ABE45: D = 0.04, p < 0.05) and one shows a bimodal trend approaching significance (ABE52: D = 0.03, p = 0.09). The remaining speakers do not exhibit bi- or multimodality (D = 0.02-0.03, p = 0.26-0.97).

Place of START: breathiness

Neither the visual inspection of the data, pooled across the speakers or by individual, nor statistical analyses reveal non-unimodality (D = 0.01-0.03, p = 0.71-0.998).

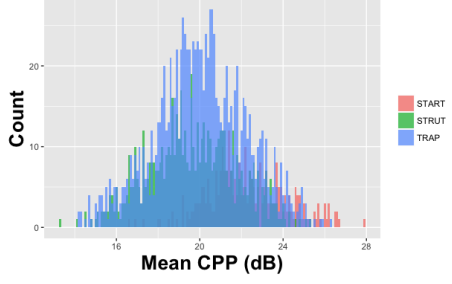

Figure 7. Distribution of mean CPP (dB) for START with respect to TRAP and STRUT.

Random forest analysis

Random forests analyses were used in order to estimate the importance of the four variables in distinguishing a. TRAP and STRUT, b. TRAP and START, and c. STRUT and START. The importance of the variables is summarised in Table 1: the higher the number, the more important the variable is in predicting the correct vowel phoneme identity (Gini importance index).

For the contrast between START and the other phonemes, vowel duration is without doubt the most important variable, following by F2 and F1 or the mean breathiness of the vowel (CPP). TRAP and STRUT, on the other hand, are again without doubt distinguished primarily by F1, following by F2, then vowel duration, and finally the mean breathiness of the vowel.

Table 1. Importance of the independent variables in predicting the three vowel contrasts; the higher the number, the more important the variable.

Pair	F1norm	F2norm	Vdur	CPP
TRAP-STRUT	424	132	85	67
STRUT-START	4	85	288	21
TRAP-START	23	87	328	16

Discussion

The analyses confirm that TRAP and STRUT are not distinguished by F2. However, the conclusion regarding F1 is more complex. On the one hand, the random forest analysis shows F1 as the most important correlate of the contrast, whose importance noticeably outperforms that of the other three variables considered. This suggests that the two phonemes have not merged regarding F1. On the other hand, the unimodality tests corroborate this result only for one of the ten speakers. Vowel duration and the mean breathiness of the vowel do not contribute to the phonetic implementation of the contrast in any obvious way, similarly to F2.

TRAP and START are distinguished primarily by vowel duration, as are STRUT and START. This is confirmed by unimodality tests for three of the ten speakers and more generally by the random forests analyses. Vowel duration emerges as the most noticeably important of the four variables considered in the implementation of the TRAP-START and STRUT-START contrasts. Although a number of studies have shown that vowel quality is essential in the perception of vowel contrasts of English, and indeed more important than vowel duration in the front part of the vocalic space at least (e.g. Bohn & Flege 1990), this does not apply to the TRAP-START and STRUT-START distinctions, at least in Aberystwyth English. Interestingly, mixed effects modelling (not shown in this paper) provides rather different results, in which all four independent variables distinguish both of the contrasts considered here.

Ultimately, perceptual experiments need to help us reach conclusions as to whether there is a TRAP-STRUT merger in Aberystwyth English.

References

Al-Tamimi, J and G Khattab (2011). Multiple cues for the singleton-geminate contrast in Lebanese Arabic: acoustic investigation of stops and fricatives. *17th ICPhS, Hong Kong*: 212-215.

Bates, D, M Maechler, B Bolker and S Walker (2015). Fitting linear mixef-effects models using lme4. *Journal of Statistical Software*, 67, 1: 1-48.

Boersma P and D Weenink (2014). Praat: doing phonetics by computer. Version 5.3.78. <http://www.praat.org/> [accessed in 2014-2015].

Bohn, O-S and J Flege (1990). Interlingual identification and the role of foreign language experience in L2 vowel perception. *Applied Psycholinguistics*, 11, 3: 303-328.

Britain, D (1992). Linguistic change in intonation: the use of high rising terminals in New Zealand English. *Language Variation and Change*, 4: 77-104.

Brunner, J and M Żygis (2011). Why do glottal stops and low vowels like each other? *17th ICPhS, Hong Kong*: 376-379.

Cox, F and S Palethorpe (2007). Australian English. *JIPA*, 37, 3: 341-350.

Crystal D (2013). *Language play*. Great Britain: Crystal ebooks.

Di Paolo M and A Faber (1990). Phonation differences and the phonetic content of the tense-lax contrast in Utah English. *Language Variation and Change*, 2: 155-204.

Eddington, D and C Channer (2010). American English has goʔ a loʔ of glottal stops: social differentiation and linguistic motivation. *American Speech*, 85, 3: 338-351.

Eddington, D and M Taylor (2009). T-glottalization in American English. *American Speech*, 84, 3: 298-314.

Fabricius, A (2007). Variation and change in the TRAP and STRUT vowels of RP: a real time comparison of five acoustic data sets. *Journal of the International Phonetic Association*, 37, 3: 293-320.

Fletcher, J, E Grabe and P Warren (2004). Intonational variation in four dialects of English: the high rising tune. *Prosodic Typology. The Phonology of Intonation and Phrasing*. Jun, S-A (ed). Oxford: OUP.

Foulkes, P and G Docherty (1999). *Urban Voices: Accents Studies in the British Isles*. London: Arnold.

Fraile R and J I Godino-Llorente (2014). Cepstral peak prominence: a comprehensive analysis. *Biomedical Signal Processing and Control* 14: 42–54.

Hall-Lew, L, R Friskney and J Scobbie (2017). Accommodation or political identity: Scottish members of the UK Parliament. *Language Variation and Change*, 29: 341-363.

Hartigan, J A and P M Hartigan (1985). The dip test of unimodality. *The Annals of Statistics*, 13, 1: 70-84.

Hejná M (2015). *Pre-aspiration in Welsh English: a case study of Aberystwyth*. PhD thesis, University of Manchester.

Hillebrand J, R A Cleveland and R L Erickson (1994). Acoustic correlates of breathy vocal quality. *Journal of Speech and Hearing Research*, 37: 769-778.

Labov W and M Baranowski (2006). 50 msec. *Language Variation and Change*, 18: 223-240.

Ligges, U and M Mächler (2003). Scatterplot3d – an R package for visualizing multivariate data. *Journal of Statistical Software*, 8, 11: 1-20.

Lotto, A J, L L Holt and K R Kluender (1997). Effect of voice quality on perceived height of English vowels. *Phonetica*, 54: 76-93.

Maechler, M (2015). Package 'diptest'. Hartigan's dip test statistic for unimodality – corrected code. <http://cran.r-project.org/web/packages/diptest/index.html> [accessed in May 2015].

MATLAB 2016b, The MathWorks, Natick (2016).

Mees, I and B Collins (1999). Cardiff: a real-time study of glottalisation. *Urban Voices: Accent Studies in the British Isles*. Foulkes, P and G Docherty (eds). London: Arnold. 185-202.

Nearey T M (1977). *Phonetic feature systems for vowels*. PhD thesis, University of Alberta.

Price. J (2008). GOOSE on the move: a study of /u/-fronting in Australian news speech. *Interspeech 2008, Brisbane*: 346.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing: Vienna, Austria. <http://www.R-project.org/> [accessed in 2014-2015].

R Studio (2009-2013). Version 0.98.1049. <http://www.rstudio.com> [accessed in 2013-2015].

Ritchart, A and A Arvaniti (2014). The form and use of uptalk in Southern Californian English. Speech Prosody 2014, Dublin.

Sando, Y T (2009). Upspeak across Canadian English accents: acoustic and sociophonetic evidence. *Proceedings of the 2009 Annual Conference of the Canadian Linguistic Association*.

Shobbrook, K (2003). High Rising Tones in Southern British English. *15th ICPhS, Barcelona*: 1273-1276.

Shokeir, V (2008). Evidence for the stable use of uptalk in South Ontario English. *NWAV*, 36, 14, 2: 15-24.

Shue Y-L (2010). The voice source in speech production: data, analysis and models. PhD thesis, UCLA.

Shue Y-L, P Keating, C Vicenik and K Yu (2011). VoiceSauce: a program for voice analysis. *17th ICPhS, Hong Kong*: 1846-1849.

Strycharczuk, P and J Scobbie (2017). Fronting of Southern British English high-back vowels in articulation and acoustics. *JASA*, 142, 1: 322.

Thomas E R and T Kendall (2007-2014). NORM: The Vowel Normalization and Plotting Suite 1.1. <http://lvc.uoregon.edu/norm/> [accessed in January 2014].

Toscano, J C and B McMurray (2010). Cue integration with categories: eighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, 34, 3: 434-464.

Ward, M (2003). Portland dialect study: the fronting of /ow, u, uw/ in Portland, Oregon. MA thesis, Portland State University.

Wells J C (1982). Accents of English 2. The British Isles. Cambridge: CUP.

Wright, M N, S Wager and P Probst (2018). Package 'range'.

Wright, M N and A Ziegler (2015). ranger: a fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77, 1: 1-17.

Deep throat as a source of information

Mattias Heldner[1], Petra Wagner[2,3] and Marcin Włodarczak[1]
[1] Department of Linguistics, Stockholm University, Sweden
[2] Faculty of Linguistics and Literary Studies, Bielefeld University, Germany
[3] Department of Speech, Music and Hearing, KTH Royal Institute of Technology, Sweden

Abstract

In this pilot study we explore the signal from an accelerometer placed on the tracheal wall (below the glottis) for obtaining robust voice quality estimates. We investigate cepstral peak prominence smooth, H1-H2 and alpha ratio for distinguishing between breathy, modal and pressed phonation across six (sustained) vowel qualities produced by four speakers and including a systematic variation of pitch. We show that throat signal spectra are unaffected by vocal tract resonances, F0 and speaker variation while retaining sensitivity to voice quality dynamics. We conclude that the throat signal is a promising tool for studying communicative functions of voice prosody in speech communication.

Introduction

Voice quality plays an important role in human communication. Voice quality contains information related to the speaker's vocal health (e.g. Maryn & Weenink, 2015; Sauder, Bretl, & Eadie, 2017). It adds to the affective content of what is being said (e.g. Airas & Alku, 2006; Gobl, 2003; Scherer, Sundberg, Tamarit, & Salomão, 2015). There are languages that employ voice quality for making phonemic contrasts (e.g. Gordon & Ladefoged, 2001; Kuang & Keating, 2014). Importantly, voice quality is also relevant for domains larger than single segments and therefore has "prosodic" functions, such as marking word- and utterance-level prominences (e.g. Kakouros, Räsänen, & Alku, 2017; Sluijter & van Heuven, 1996; Yanushevskaya, Chasaide, & Gobl, 2017), boundary phenomena (e.g. Carlson, Hirschberg, & Swerts, 2005) and turn-taking (e.g. Gravano & Hirschberg, 2011; Ogden, 2002). Such observations of suprasegmental functions of voice quality led to coining the term *voice prosody* for studies of the communicative functions of voice (Gobl, Yanushevskaya, & Chasaide, 2015).

Consequently, voice quality dynamics is a relevant topic in speech communication research. A major obstacle in this pursuit, however, is that it is difficult to measure relevant aspects of voice quality in a reliable way, and especially in continuous or conversational speech. There are several reasons for this. First of all, a lot of voice quality research is based on sustained vowels and many established voice quality measures (e.g. jitter and shimmer) require such data in order to be meaningful. Furthermore, voice quality measurements often involve *glottal inverse filtering* techniques to remove the effects of the vocal tract and the lip radiation from the microphone signal (e.g. effects of formants on spectrum slope). While automatic inverse filtering techniques exist (see e.g. Alku, 2011 for a review), they are generally not considered accurate enough when applied to continuous speech (Alku, 2011; Gobl, et al., 2015). Thus, voice quality researchers often resort to manual glottal inverse filtering, which is both very time consuming and requires highly skilled experimenters (Gobl, et al., 2015). As a consequence, voice quality studies on large-scale conversational speech are scarce.

Inspired by recent work using accelerometers (aka throat microphones) placed on the neck surface below the glottis for ambulatory voice monitoring (Mehta, et al., 2015), as well as own recent experiences with throat microphones for capturing breathing noises (Włodarczak & Heldner, 2017), in this pilot study we explore accelerometer signals for obtaining voice quality measures. Thus, the primary goal of this paper is to explore whether accelerometers placed on the tracheal wall are sensitive to voice quality dynamics without the need for inverse filtering of the throat microphone signal (as in Chien, Mehta, Guenason, Zañartu, & Quatieri, 2017; Llico, et al., 2015; Zañartu, Ho, Mehta, Hillman, & Wodicka, 2013). A secondary goal is to evaluate the robustness of the throat microphone signal to formant variation, pitch variation, and pitch level. While the long-term goal is applying such methods to continuous speech, we take sustained vowels as a starting point here.

Materials & methods

Subjects

Three semiprofessional singers (2 females, 1 male) with phonetic expertise and one expert phonetician (1 male) served as voice talents.

Recording

All recordings took place in a sound treated room at Stockholm University. During recording, participants produced sequences of 6 sustained vowels /aː, eː, iː, yː, uː, oː/ at 4 different pitch levels each, covering one octave, and with 3 different voice qualities (modal, breathy, tense). The recordings were ordered by voice quality, that is participants chose an individually comfortable low pitch level, a vowel and a voice quality to start with, e.g. modal /aː/, and then produced modally voiced sequences for each vowel starting from their low comfort pitch level, then successively raising pitch by a major third until a full octave was reached, and then successively lowering pitch until the base pitch is reached again. The participants produced the same sequence for the remaining voice qualities. Participants were asked to target 1-2 seconds for each sustained vowel, but neither durations nor pitch levels were strictly controlled. Per speaker, each combination of *vowel-pitch-quality* is recorded twice, except for the highest pitch level, and 7 recordings were made for each *vowel-quality* combination. In total, 672 vowels were recorded.

Data acquisition

The speech signal was recorded using a directional headset microphone (DPA 4088) placed 3 cm from the corner of the mouth. This microphone has a flat frequency response up to 1 kHz and a soft boost (4-6 dB) up to 15 kHz. The throat signal was recorded using a miniature accelerometer (Knowles BU-27135) attached to the skin on the tracheal wall (below the cricoid cartilage) with cosmetic glue (see Figure 1). This accelerometer has a flat frequency response from 20 Hz to 3 kHz and a 4 dB boost up to 6 kHz. We use the same accelerometer as in Mehta, et al. (2015), and the sensor was made in the Phonetics Laboratory at Stockholm University.

Figure 1. Accelerometer attached to the skin on the tracheal wall.

Both signals were connected to a Shure ULX-D digital wireless system and recorded using the REAPER software.

Acoustic measures

We captured three aspects of voice quality: (i) signal periodicity, (ii) the relative amplitude of the first harmonic, and (iii) spectral tilt.

The signal periodicity was assessed by Cepstral Peak Prominence (CPP, Hillenbrand, Cleveland, & Erickson, 1994). Defined as the amplitude of the first peak in the real cepstrum (first rahmonic) of a sound, relative to the cepstrum trend line, CPP has been used extensively in clinical literature as a measure of dysphonia (Sauder, et al., 2017). It also been successfully used for detection of breathiness and, with somewhat mixed results, for assessment of the overall voice quality (see Fraile & Godino-Llorente, 2014 for a review). In this paper, we used the smoothed version of CPP (CPPS, Hillenbrand & Houde, 1996), following the procedure outlined in Watts, Awan, and Maryn (2017).

The relative amplitude of the first harmonic (i.e. the fundamental) was measured using H1-H2, which is a measure of the amplitude of the first harmonic in dB relative to the second harmonic (Hillenbrand & Houde, 1996). Note however, that H1-H2 can also be viewed as a measure of spectral tilt (in dB per octave) in the lower part of the spectrum (cf. Kakouros, et al., 2017; Titze & Sundberg, 1992).

Spectral tilt was measured using the alpha ratio (Frokjaer-Jensen & Prytz, 1976), which is a measure of spectral balance, defined as a ratio of energy below and above 1000 Hz.

Each of the measures was calculated for the speech signal as well as for the throat signal. H1-

H2 was additionally calculated for an estimation of the voice source in the speech signal obtained using an automatic inverse filtering method (Airas, 2008; Alku, 1992). All measures were *z*–normalized by speaker and microphone.

We have used freely available speech processing tools for all of the analyses in this paper. CPPS and alpha ratio were calculated in Praat (Boersma & Weenink, 2018), H1-H2 was obtained from the COVAREP repository (Degottex, Kane, Drugman, Raitio, & Scherer, 2014). All features were subsequently speaker-normalized. Additionally, since H1-H2 is likely to be affected by speaker's F0, we split values of these features of the median F0 calculated for all speakers (186 Hz).

Analyses

For this pilot study, we restricted the analyses to (i) qualitative descriptions or illustrations of why the throat signal may provide a more robust estimation of voice quality and (ii) descriptive statistics of the acoustic measures to allow a comparison of how well the different measures separate the different voice qualities.

Results

A first illustration of why the throat signal may potentially be useful for voice quality measures is given in Figure 2 showing LPC spectra of the same vowel from a microphone signal, a throat signal and an inverse filtered microphone signal using an LPC based inverse filter function in Praat. It is easy to see that vocal tract formants will influence any microphone-based characterization of spectral tilt involving the F1 to F3 frequency region. In contrast, there is no evident influence of vowel formants in the throat signal, although resonances that most likely originate from the subglottal system are visible at approximately 550, 1400 and 2700 Hz (cf. Sundberg, Scherer, Hess, Muller, & Granqvist, 2013).

It is also evident from Figure 2 that the throat signal spectrum is different from the voice source spectrum estimated using inverse filtering. In particular, the throat signal spectrum has an elbow at the first subglottal resonance, whereas the inverse-filtered signal rolls off monotonously. Thus, the throat spectrum is perhaps better characterized by a two-segment slope below and above the first subglottal resonance, or by a polynomial function.

A second, and perhaps more convincing illustration of the benefits of the throat signal, is provided in Figure 3 showing LPC spectra of the throat signals for three different vowels by the same speaker. The three spectra are virtually identical. This indicates that the throat signal is robust to variations in formant frequencies. Similar analyses with different speakers (Figure 4) and with different F0 levels (Figure 5) show that the throat signal is also robust to speaker and pitch variation.

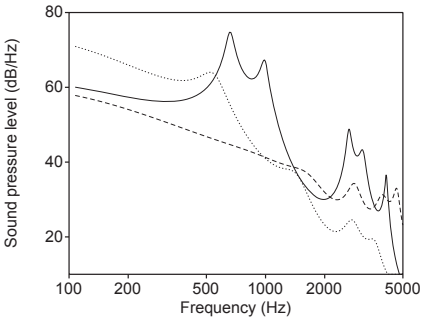

Figure 2. LPC spectra of the vowel /aː/ *from the normal microphone signal (solid line), the throat signal (dotted line), and an inverse filtered microphone signal (dashed line). The vowel was produced in modal voice quality by a male speaker (F0 ≈ 115 Hz).*

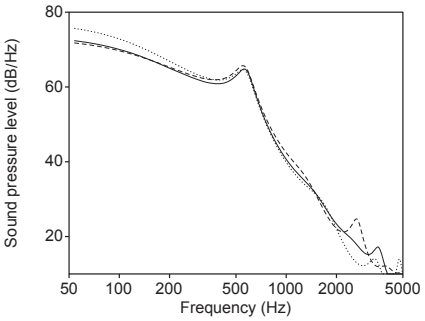

Figure 3. LPC spectra of three different vowels (/aː/ *solid line,* /iː/ *dotted line,* /uː/ *dashed line) produced in modal voice quality by a male speaker (F0 ≈ 150 Hz).*

But of course, it is not enough for a signal to be robust to various influences in order to be useful for voice quality measures. It has to be sensitive to relevant voice quality variation as well. Figure 6 illustrates this aspect of throat signals with LPC spectra of different voice qualities (same vowel, same speaker).

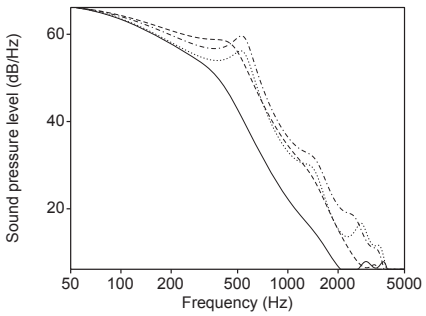

Figure 4. LPC spectra of modal voice /a:/ by four speakers (2f, 2m). For comparison, the spectra have been shifted on the y-axis.

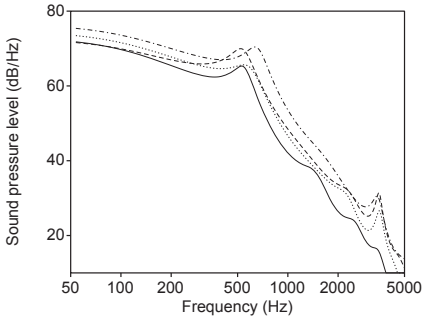

Figure 5. LPC spectra of the vowel /a:/ at four different F0 levels. Female speaker.

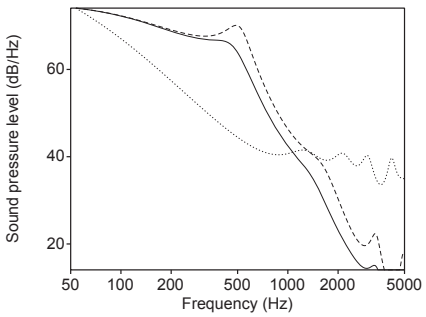

Figure 6. LPC spectra of three voice qualities (modal: solid line, tense: dotted line, breathy: dashed line) produced on the same vowel /a:/ *by a female speaker (F0 ≈ 150 Hz). For comparison, the spectra have been shifted on the y-axis.*

These conclusions are further confirmed by the results in Figures 7–9, which show that the throat signal provides a robust separation between the three voice qualities regardless of the measure used. In Figure 7, we plot CPPS values calculated from the throat and speech signals. Overall, the signal periodicity increases from breathy to modal to pressed. Not surprisingly given its original purpose, CPPS is very effective at distinguishing breathy and non-breathy phonations in both types of signals. Additionally, in the throat signal it also offers better separation of modal and pressed phonations.

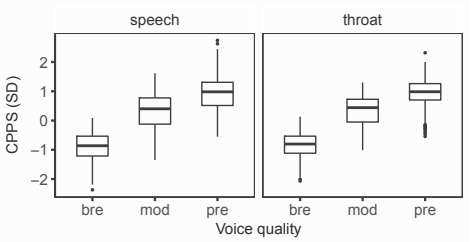

Figure 7. Box plots of normalized signal periodicity (CPPS) in breathy (bre), modal (mod) and pressed (pre) voice quality.

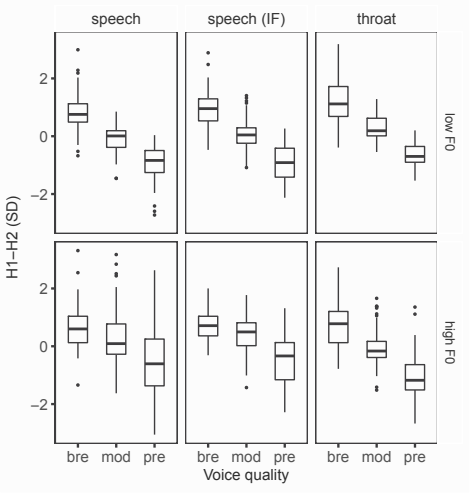

Figure 8. Box plots of normalized relative amplitude of the first harmonic (H1-H2) in breathy (bre), modal (mod) and pressed (pre) voice quality. Data is split by F0 level with the values below median in the top row and above median in the bottom row.

H1-H2 (Figure 8), reflecting both relative amplitude of the first harmonic and spectral tilt in the lower part of the spectrum, shows large dependence on F0 level when calculated on the speech signal. Namely, it separates the three voice qualities rather well at low F0 levels but fails for higher F0 values (especially for the breathy-modal contrast). This is most likely due to the fact that for higher pitches the first two har-

monics are increasingly influenced by F1. Notably, this effect is also observed in the automatically inverse-filtered signal, suggesting that residuals of vocal tract resonances must be present in the signal. By contrast, the throat signal is virtually unaffected by fundamental frequency.

Finally, spectral balance captured by alpha ratio (Figure 9) has little discriminatory value when calculated on the speech signal but preserves good separation in the throat signal.

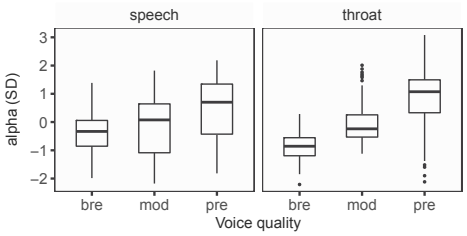

Figure 9. Box plots of normalized spectral tilt (alpha ratio) in in breathy (bre), modal (mod) and pressed (pre) voice quality.

Discussion

Based on the results in the previous section, we conclude that the throat signal is robust to vowel quality, speaker and pitch variation. At the same time, it is sensitive to differences in voice quality. This is the case even though the throat signal is spectrally different from signals obtained using inverse filtering of speech signals (prevalent in voice quality research). Given its robustness and stability, we speculate that we can also eliminate the need for inverse filtering of the throat signal (e.g. Zañartu, et al., 2013) for the purpose of studying the communicative function of voice quality dynamics.

In future work, we will investigate whether the the throat signal is equally useful for studies of continuous, spontaneous or conversational speech. We will also monitor sound pressure level (SPL) given the known dependency between SPL and spectral tilt (e.g. Sundberg & Nordenberg, 2006). Finally, we are planning to evaluate other voice quality measures. In particular, we hope to obtain a better estimate of spectral tilt by using a DNN-based approach (Jokinen & Alku, 2017; Kakouros, et al., 2017), and to explore measures of pitch-strength (Eddins, Anand, Camacho, & Shrivastav, 2016) as an alternative to H1-H2.

In conclusion, the throat signal is a promising, tool for studying communicative functions of voice prosody in speech communication. It could potentially allow quantitative analyses of large speech materials without relying on the error prone automatic inverse filtering methods of speech signals.

Acknowledgements

This work was partly funded by a Stiftelsen Marcus och Amalia Wallenbergs Minnesfond project MAW 2017.0034 *Hidden events in turn-taking* to the first author; by a Humbolt stipend within the Swedish-German Programme *Research Awards for Scientific Cooperation* to the second author; and by a Christian Benoît Award to the third author.

References

Airas M (2008) TKK Aparat: an environment for voice inverse filtering and parameterization. *Logopedics Phoniatrics Vocology* 33: 49-64. doi: 10.1080/14015430701855333.

Airas M and Alku P (2006) Emotions in vowel segments of continuous speech: analysis of the glottal flow using the normalised amplitude quotient. *Phonetica* 63: 26-46. doi: 10.1159/000091405.

Alku P (1992) Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering. *Speech Communication* 11: 109-118. doi: 10.1016/0167-6393(92)90005-r.

Alku P (2011) Glottal inverse filtering analysis of human voice production — A review of estimation and parameterization methods of the glottal excitation and their applications. *Sadhana* 36: 623-650. doi: 10.1007/s12046-011-0041-5.

Boersma P and Weenink D. (2018). Praat: doing phonetics by computer [Computer program] (Version 6.0.39). Retrieved from http://www.praat.org/

Carlson R, Hirschberg J and Swerts M (2005) Cues to upcoming Swedish prosodic boundaries: Subjective judgment studies and acoustic correlates. *Speech Communication* 46: 326-333. doi: 10.1016/j.specom.2005.02.013.

Chien Y-R, Mehta D D, Guenason J, Zañartu M and Quatieri T F (2017) Evaluation of Glottal Inverse Filtering Algorithms Using a Physiologically Based Articulatory Speech Synthesizer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25: 1718-1730. doi: 10.1109/taslp.2017.2714839.

Degottex G, Kane J, Drugman T, Raitio T and Scherer S. (2014). COVAREP - A collaborative voice analysis repository for speech technologies *Proc. ICASSP 2014* (pp. 960-964). Florence, Italy.

Eddins D A, Anand S, Camacho A and Shrivastav R (2016) Modeling of breathy voice quality using pitch-strength estimates. *Journal of Voice* 30: 774 e771-774 e777. doi: 10.1016/j.jvoice.2015.11.016.

Fraile R and Godino-Llorente J I (2014) Cepstral peak prominence: A comprehensive analysis. *Biomedical*

Signal Processing and Control 14: 42-54. doi: 10.1016/j.bspc.2014.07.001.

Frokjaer-Jensen B and Prytz S (1976) Registration of voice quality. *Brüel and Kjær Technical Review* 3: 3-17.

Gobl C (2003) The role of voice quality in communicating emotion, mood and attitude. *Speech Communication* 40: 189-212. doi: 10.1016/s0167-6393(02)00082-1.

Gobl C, Yanushevskaya I and Chasaide A N (2015) The relationship between voice source parameters and the maxima dispersion quotient (MDQ). In *Proc. Interspeech 2015*. Dresden, Germany, 2337-2341.

Gordon M and Ladefoged P (2001) Phonation types: a cross-linguistic overview. *Journal of Phonetics* 29: 383-406. doi: 10.1006/jpho.2001.0147.

Gravano A and Hirschberg J (2011) Turn-taking cues in task-oriented dialogue. *Computer Speech & Language* 25: 601-634. doi: 10.1016/j.csl.2010.10.003.

Hillenbrand J, Cleveland R A and Erickson R L (1994) Acoustic Correlates of Breathy Vocal Quality. *Journal of Speech Language and Hearing Research* 37. doi: 10.1044/jshr.3704.769.

Hillenbrand J and Houde R A (1996) Acoustic Correlates of Breathy Vocal Quality: Dysphonic Voices and Continuous Speech. *Journal of Speech Language and Hearing Research* 39. doi: 10.1044/jshr.3902.311.

Jokinen E and Alku P (2017) Estimating the spectral tilt of the glottal source from telephone speech using a deep neural network. *Journal of the Acoustical Society of America* 141: EL327. doi: 10.1121/1.4979162.

Kakouros S, Räsänen O and Alku P (2017) Evaluation of spectral tilt measures for sentence prominence under different noise conditions. In *Proc. Interspeech 2017*. Stockholm, Sweden: ISCA, 3211-3215. doi: 10.21437/Interspeech.2017-1237.

Kuang J and Keating P (2014) Vocal fold vibratory patterns in tense versus lax phonation contrasts. *Journal of the Acoustical Society of America* 136: 2784-2797. doi: 10.1121/1.4896462.

Llico A F, Zañartu M, Gonzalez A J, Wodicka G R, Mehta D D, Van Stan J H, et al. (2015) Real-time estimation of aerodynamic features for ambulatory voice biofeedback. *Journal of the Acoustical Society of America* 138: EL14-19. doi: 10.1121/1.4922364.

Maryn Y and Weenink D (2015) Objective dysphonia measures in the program Praat: smoothed cepstral peak prominence and acoustic voice quality index. *Journal of Voice* 29: 35-43. doi: 10.1016/j.jvoice.2014.06.015.

Mehta D D, Van Stan J H, Zañartu M, Ghassemi M, Guttag J V, Espinoza V M, et al. (2015) Using ambulatory voice monitoring to investigate common voice disorders: Research update. *Frontiers in Bioengineering and Biotechnology* 3: 155. doi: 10.3389/fbioe.2015.00155.

Ogden R (2002) Turn transition, creak and glottal stop in Finnish talk-in-interaction. *Journal of the International Phonetic Association* 31. doi: 10.1017/s0025100301001116.

Sauder C, Bretl M and Eadie T (2017) Predicting voice disorder status from smoothed measures of cepstral peak prominence using Praat and Analysis of Dysphonia in Speech and Voice (ADSV). *Journal of Voice* 31: 557-566. doi: 10.1016/j.jvoice.2017.01.006.

Scherer K R, Sundberg J, Tamarit L and Salomão G L (2015) Comparing the acoustic expression of emotion in the speaking and the singing voice. *Computer Speech & Language* 29: 218-235. doi: 10.1016/j.csl.2013.10.002.

Sluijter A M C and van Heuven V J (1996) Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America* 100: 2471-2485. doi: 10.1121/1.417955.

Sundberg J and Nordenberg M (2006) Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech. *The Journal of the Acoustical Society of America* 120: 453-457. doi: 10.1121/1.2208451.

Sundberg J, Scherer R, Hess M, Muller F and Granqvist S (2013) Subglottal pressure oscillations accompanying phonation. *Journal of Voice* 27: 411-421. doi: 10.1016/j.jvoice.2013.03.006.

Titze I R and Sundberg J (1992) Vocal intensity in speakers and singers. *The Journal of the Acoustical Society of America* 91: 2936-2946. doi: 10.1121/1.402929.

Watts C R, Awan S N and Maryn Y (2017) A comparison of cepstral peak prominence measures from two acoustic analysis programs. *Journal of Voice* 31: 387 e381-387 e310. doi: 10.1016/j.jvoice.2016.09.012.

Włodarczak M and Heldner M (2017) Capturing respiratory sounds with throat microphones. In: J Eggesbø Abrahamsen, J Koreman & W A van Dommelen, eds, *Nordic Prosody: Proceedings of the XIIth Conference, Trondheim 2016*. Frankfurt am Main, Germany: Peter Lang, 181-190.

Yanushevskaya I, Chasaide A N and Gobl C (2017) Cross-speaker variation in voice source correlates of focus and deaccentuation. In *Proc. Interspeech 2017*. Stockholm, Sweden: ISCA, 1034-1038. doi: 10.21437/Interspeech.2017-1535.

Zañartu M, Ho J C, Mehta D D, Hillman R E and Wodicka G R (2013) Subglottal Impedance-Based Inverse Filtering of Voiced Sounds Using Neck Surface Acceleration. *IEEE Transacions on Audio, Speech, and Language Processing* 21: 1929-1939. doi: 10.1109/TASL.2013.2263138.

Intelligibility of the alveolar [s] replacing the initial interdental /θ/ in English words

Hyeseung Jeong and Bosse Thorén
Department of Social and Behavioural Studies, University West, Sweden

Abstract

The study examines the intelligibility of a German speaker's replacement of the initial interdental /θ/ with the alveolar fricative [s] in words that occurred in her reading of a short English text. Twenty nine students in university English courses in Sweden listened to, and transcribed the whole reading, where substituting the initial /θ/ of a word with [s]appeared four times. The result shows that the phoneme substitution by the German speaker did not cause misunderstanding in three instances, but it considerably misled the listeners' understanding of a phrase in one occasion. We discuss this finding in relation to the functional load of the initial θ/s contrast (Catford, 1987), and Jenkins' (2002, 2015) Lingua Franca Core syllabus.

Introduction

The interdental fricatives /ð/ and /θ/ are not easy to articulate, and they are less frequent than other phonemes in the world's languages (Jenkins, 2015). English is one of the languages that have the two interdentals, but replacing them with some other sounds is a common phenomenon among many of its speakers (Jenkins, 2000; Kirkpatrick, 2010; Pennington, 1996). For example, the voiceless /θ/, which this study is concerned with, is substituted with the alveolar /t/ and the post-dental /t̪/ 'in many areas of Britain and in many indigenous varieties of English, such as African and Caribbean, as well as in many learner varieties' (Jenkins, 2000, p 137). The labiodental /f/ and the alveolar /s/ are also often found as what replace the /θ/, while substituting with /f/ is usually done by L1 speakers and with /s/ mostly by L2 speakers, such as Japanese or German speakers.

In addition to reporting its commonness, research also suggests that the substitution of /θ/ (and /ð/) may not be one of the phonetic features that seriously threaten the intelligibility of a person's English pronunciation. For example, earlier, Catford (1987) introduced the relative functional load of different English phoneme contrasts, which says that the phoneme pairs with high functional load (e.g., the initial /p/ and /b/) hurt intelligibility more than those with low functional load (e.g., the initial /s/and /z) when they are not properly contrasted. Later Munro and Derwing (2006) validated this notion of Catford's through empirical studies. According to Catford's relative functional load table represented by Derwing and Munro (2014, p. 49), the initial/θ/ and /s/ contrast has relatively low functional load at 21 %. In addition, Jenkins (2000, 2002) reported that pronouncing /θ/ as /t/ or /s/ did not cause miscommunication among international interlocutors. Based on this finding, she classified /θ/ as a non-core feature in her Lingua Franca Core syllabus (Jenkins, 2000, 2015).

Against this backdrop, the study examines the intelligibility of a German speaker's alveolar fricative [s] that replaced the initial interdental fricative /θ/ in English words, based on the perceptions of university students in English courses in Sweden. We want to see whether, and to what extent the replacement of /θ/ with [s] in different contexts conforms to what the literature tells about such phoneme substitution.

Method

Subjects

The subjects were twenty nine people that enrolled in two different English courses at a university in Sweden. Twenty six of them were Swedish L1 speakers and three were English L1speakers.

Material

We used the audio-recording of a German speaker reading a short text in English. The recording was from the Speech Accent Archive, a free online resource (accent.gmu.edu), provided by the linguistic program at George Mason

University in the United States. The whole text is as follows, where the initial voiceless interdental fricative /θ/ in four words are underlined:

'Please call Stella. Ask her bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three bags, and we will go meet her Wednesday at the train station.'

The German speaker pronounced the subjected phoneme in the four words as the voiceless alveolar fricative [s]. Table 1 presents the IPA transcript of the speaker for the four words starting with the phoneme, together with the words appearing before and after them, to understand the contexts that influenced the subjects' perceptions. It also provides the IPA transcripts of Received Pronunciation (RP) and General American (GA) pronunciation for the same word strings for comparison.

Table 1.the German speaker' initial alvelolars that replaced the initial interdental in RP and GA.

word strings	German speaker	RP	GA
these things with	d̪iːs s̪ɪŋks wɪθ	/ðiːz θ̲ɪŋz wɪð/	/ðiz θ̲ɪŋz wɪð/
five thick slabs	faɪf s̪ik slæbs	/faɪv θ̲ɪk slæbz/	/faɪv θ̲ɪk slæbz/
these things into	d̪is s̪iŋs ɪntu	/ðiːz θ̲ɪŋz ɪntə/	/ðiz θ̲ɪŋz ɪntə/
into three red	ɪntu s̪ʁi ʁet	/ɪntə θ̲riː rɛd/	/ɪntə θ̲ri rɛd/

Transcription task

As part of their course assignement, the subjects transcribed the speaker individually. They could listen to the speaker as many times as they wanted. However, preciseness in their transcription was not what they were asked to achieve. They were instead told to transcribe as they heard and understood. The subjects knew that the given task was checking the degree of the intelligibility of the speaker, rather than making an errorless transcription.

Result

Table 2 shows that the replacement of the initial voiceless interdental fricative with the voiceless alveolar fricative by the German speaker greatly hurt the intelligibility of her pronunciation in one of the four occasions. In the other three, the subjects had very little or no problem with understanding the speaker.

Table 2. Error rate in the subjects' perception on the German speaker' initial alvelolars that replaced the initial interdental in RP and GA.

word strings	the German speaker's realisation	error rate in the perceptions of the subjects* (%)
these things with	d̪iːs s̪ɪŋks wɪθ	0
five thick slabs	faɪf s̪ik slæbs	89.6
these things into	d̪is s̪iŋs ɪntu	0
into three red	ɪntu s̪ʁi ʁet	0

*The number of the subjects = 29

As seen in the table, no one mistook the interdental /θ/ 'in these things with', 'these things into' and 'into three red' for some other sounds. An English L1 listener thought 'these things with' as '*this thing* with', and a Swedish L1 listner transcribed the same phrases as '*this* things with' and 'these things into' as '*this* things into'. This, however, seemed to be due to the fact that replacing the final voiced /z/ with the voiceless [s] 'these' /ðiːz. Except for the two listeners, all other listeners correctly transcribed the two phrases, and 'into three red' was precisely perceived by all the subjects.

By contrast, only three out of twenty six Swedish L1 listeners and 1 out of three English L1 listeners correctly transcribed the word 'thick' in 'five thick slabs', and all the others perceived it as 'six'. There were some variants among those who mistook the word for 'six' in their transcriptions. A Swedish L1 subject perceived

the phrase with 'thick' as 'five six slash', another Sweidsh L1 subject as 'five six slap', three Swedish L1 subjects as 'five six laps', and an English L1 subject as 'five or six slabs'. Otherwise the phrase was transcribed as 'five six slabs'.

There can be two possible reasons for this high rate of misunderstanding. First, the sequence of the first four sound segments in the German speaker's pronunciation of 'thick slabs' [sik slæbs] is exactly how the word 'six' sounds: /siks/. Second, just before the sound sequence, the word 'five' appears, which was correctly perceived by all the subjects, although its final consonant /v/ was replaced with /f/ by the speaker. Consequently, 'five, six' as two alternatives for the number of cheese slabs seemed to emerge in the listeners' minds as a reasonable interpretation for what they could not clearly hear.

Discussion and Conclusion

The finding shows that only one out of the four instances of replacing the word initial /θ/ with [s] by a German speaker caused misunderstanding, while the other three did not. This is, on the whole, 22.4% of error rate when 26 errors were divided by 116 observations (4 words x 29 subjects). The result does not challenge what is in the literature. The error rate is just a little higher than the 21 % functional load of the initial θ/s contrast in the table by Catford (Derwing & Munro, 2014, p. 49). Given that the subjects of the study correctedly understood the messages in the three substituting instances, the result does not really discord with Jenkin's Lingua Franca Core syllabus that classifies the initial /θ/ as non-core.

Nevertheless, the finding clearly indicates that in certain contexts pronouncing /θ/ as /s/ can still decrease phonetic intelligibility, even to a great exent. This possibility was once discussed by Brown (1974 cited in Jenkins, 2000), who therefore suggested that if an L2 speaker had difficulty with realising the voiceless interdental, she/he could replace with the labiodental fricative /f/ rather than the alveorlar fricative. However, the functional load of the contrast between /θ/ and /f/ contrast is 15%, and between /θ/ and /t/ (another phoneme that frequently replaces /θ/) is 18 %, both of which are just slightly lower than the 21 % load of the /θ/ and /s/ contrast. In addition, if 'thick' in 'five thick slabs' were pronounced as 'fick' or 'tick', that may cause less confusion for the listeners. However, such replacements may cause problems in other contexts: for example, pronouncing the phrase, 'into three red (bags)' as 'into free red (bags)' or 'into tree red (bags)' can create a misunderstanding, which did not arise when the phrase was pronounced as 'into sri red' Therefore, it may not be so helpful to encourage learners to replace /θ/ with /t/ or /f/ instead of /s/ to prevent miscommunication. Instead, based on our finding, we rather wish to suggest that English teachers help learners to be aware that not properly realising /θ/ can be problematic sometimes, although they may not need to spend too much time on learning the phoneme.

References

Catford, J C (1987). Phonetics and the teaching of pronunciation: A systemic description of English phonology. In J Morley (Ed.), *Current perspectives on pronunciation: Practices anchored in theory* (pp. 87-100). Alexandria, VA: TESOL.

Derwing, T & Munro, M (2014). Once you have been speaking a second language for years, it's too late to change your pronunciation. In L Grant (Ed.), *Pronunciation myths: Applying second language research to classroom teaching* (pp. 34-57). Ann Arbor: University of Michigan Press.

Jenkins, J (2000). *The phonology of English as an international language.* Oxford, England: Oxford University Press.

Jenkins, J (2002). A sociolinguistically based, empirically researched pronunciation syllabus for english as an international language. *Applied Linguistics*, 23(1), pp. 83-103. doi:10.1093/applin/23.1.83

Jenkins, J (2015). *Global Englishes: A resource book for students.* London, England: Routledge.

Kirkpatrick, A (Ed.). (2010). *The Routledge handbook of World Englishes*. London, England: Routledge.

Munro, M & Derwing, T (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, 34(4), pp. 520-531. doi.org/10.1016/j.system.2006.09.004

Pennington, M (1996). *Phonology in English language teaching: An international approach.* London: Longman.

The perceptual effect of voicedness in laughter

Kristina Lundholm Fors and Ellen Breitholtz.
Språkbanken, Department of Swedish & Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg

Abstract

Laughter can be used to express a range of emotions. We may differentiate between cooperative laughter – expressing for example enjoyment or agreement – and non-cooperative laughter, expressing for example mocking or disagreement (Tian et al., 2016). Our aim in this study was to investigate how voicedness influences the perception of the pragmatic function of laughter. Specifically, we were interested in the type of laughter called "hånskratt" in Swedish, which may be translated as mocking or jeering laughter. To explore this, we performed a pilot study in the form of a perception experiment.

Humans are adept at distinguishing between positive and negative laughter (Devillers & Vidrascu, 2007), and voicing may be a relevant factor in determining whether a laughter event is mocking or not (Bachorowski & Owren, 2001). According to Laskowski & Burger (2007), voicing is also connected to the situational context, in that voiced, unlike un-voiced, laughter is more likely to be preceded and succeeded by the laughter of others. Our hypothesis was therefore that unvoiced laughter would be more likely to be perceived as mocking, but that by adding pragmatic context, this interpretation could be modified.

For the perception experiment, four two-part image series were used, depicting stick figures in the following situations: 1) one person falling over, and laughing about it, 2) two persons falling over, and laughing about it, 3) one person watching a clown on tv, and laughing, and 4) one person falling over, and another person pointing and laughing at that person, while the person who fell over looks sad. With each set of images, the study participant heard a laughter sample (voiced laughter/unvoiced laughter/impression of mocking laughter) and were told to click on the image they thought best corresponded with the laughter.

Results showed that voicedness was not primarily associated with mocking laughter, but rather with the person being alone, whereas voiced laughter was associated with two persons laughing together. Mocking laughter was linked to the person laughing at someone falling over. Adding the utterance "de va rätt åt dig" ("that serves you right") lead to subjects associating the voiced and unvoiced laughter samples with the image of the person being laughed at for falling over.

From this pilot study we draw the tentative conclusion that voiced and unvoiced laughter are associated with different types of laughter-inducing situations. However, unvoiced laughter was not identified primarily as mocking laughter, but rather seems to be perceived as the laughter of someone who is laughing by themself. Further, our results indicate that the perceived meaning of laughter can be modified by the context in which the laughter appears.

References

Bachorowski JA & Owren MJ (2001). Not All Laughs are Alike: Voiced but Not Unvoiced Laughter Readily Elicits Positive Affect. *Psychological Science*, *12*(3), 252–257.

Devillers L & Vidrascu L (2007). Positive and negative emotional states behind the laughs in spontaneous spoken dialogs. In *Interdisciplinary Workshop on The Phonetics of Laughter*: 37– 40.

Laskowski K & Burger S (2007). On the correlation between perceptual and contextual aspects of laughter in meetings. In *Interdisciplinary Workshop on Phonetics of Laughter*: 55–60.

Tian Y, Mazzocconi C & Ginzburg J (2016). When do we laugh? In *Proceedings of the SIGDIAL 2016 Conference*: 360–369. Association for Computational Linguistics.

Perception and Production of L2 prosody by Swedish Learners – Summary and Application to the teaching using Japanese and Chinese Data

Yasuko Nagano-Madsen
Department of Langauges and Literatures, University of Gothenburg

Abstract

The present paper is about the perception and production of L2 prosody by Swedish learners with focus on L2 Chinese and Japanese by the author. It gives the summary of the author's previous and ongoing studies on the topic as well as discussion and application to teaching. Swedish L2 prosody is characterized by the upward pitch movement, i.e. preference of F0 rise to F0 fall, preference of upstep to downstep, and lack of a sharp F0 fall in their production. Furthermore, the Swedish learners have a difficulty in distinguishing F0 rise from F0 fall in perception.

Introduction

In this paper, some of the previous studies on the perception and production of L2 prosody by Swedish learners with focus on L2 Chinese and Japanese are summarised. It also presents the discussion and application to the teaching that are based on the results.

Although studies of L2 speech acquisition and processing are many, studies that involve either a pitch-accent language or a tone language as source and target languages are still limited. There are three main language groups that differ in the use of pitch for linguistic purpose. First is the 'true' tone languages such as Chinese, Thai, Vietnamese, and Yoruba. In these languages, tones have to be introduced right from the beginning together with the lexicon since tone is an indispensable part of the lexicon. The second is the lexical pitch accent languages such as Japanese and Swedish in which pitch accent does not play as important role as tone does in a tone language. In these languages, pitch accent realization varies greatly among the dialects, yet speakers can understand each other without much difficulty. In addition, it is common for a pitch accent language to have an area in which the accentual contrast are lost. In the third group, which contains most of the European languages including English and German, the function of pitch accent is only 'post lexical', i.e. it is important for intonation. The large majority of languages in Africa and Asia are either lexical pitch-accent languages or tone languages. Swedish, Norwegian and Serbo-Croatian are the few lexical pitch-accent languages found in Europe.

Swedish

Figure 1 below shows the F0 contours of two types of pitch accent in Swedish, *anden* 'the duck' (accent 1) and *anden* 'the spirit' (accent 2). Both accent 1 and accent 2 have a pitch rise and a fall, but they differ in the timing of F0 movements with the segments. The critical manifestation of the lexical pitch accents in Swedish is the timing of F0 gesture with segments while pitch register is only relevant for discourse function, i.e. for the manifestation of focus (Gårding 1973, Bruce 1977). The phonological difference between the Swedish pitch accent and Japanese pitch accent are discussed in Nagano-Madsen and Bruce (1998).

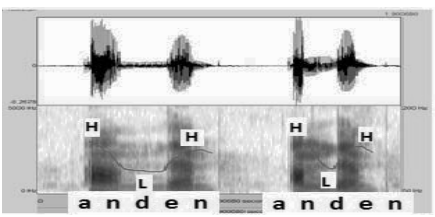

Figure 1. The two types of lexical pitch accents in Swedish. West coast dialect, a male speaker.

Mandarin Chinese

Word level prosody

Mandarin Chinese (henceforth Chinese) has four lexical tones. The F0 configurations for /mā, má, mǎ, mà/ spoken by a native female speaker are shown in Figure 2. Note that the four tones differ not only in the pitch patterns but also in other features. T4 is the shortest while the duration of T3 is twice as long as that of T4. In addition, T4 regularly contained a creaky voice (dotted line in the figure) at a low pitch region.

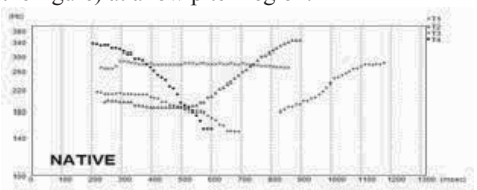

Figure 2. F0 configurations for /mā, má, mǎ, mà/ in Mandrin. Female speaker.

Perception of Mandarin tones

Nagano-Madsen and Won (2016) examined the perception of Mandarin tones in mono- and di-syllabic words by Swedish learners of Chinese at three proficiency levels. It reveals that the L2 acquisition process of tonal perception is not a quantitative change but also a qualitative one. We can speak of 'interlanguage' here, i.e. a unique linguistic system that changes continuously during the course of acquisition (Selinker, 1972). As for the monosyllabic words, the correct identification score was in the order of T3 > T1 > T4 > T2 for the beginner level. The effect of glotalization in identifying T3 was found to override a pitch cue for monosyllabic words but not for di-syllabic words.

The major difficulties in tonal perception for monosyllabic words disappear by the time of intermediate level (18 months) except for the confusion between Tone 2 (T2) and Tone 3 (T3). Confusion between a falling tone (T4) with a rising tone (T2 or T3) was common at the beginner level but not at the intermediate level. The results of our study also show that the tonal perception of Mandarin tones differs greatly depending on whether the word in question is monosyllabic or disyllabic, and for the latter, whether it is the first syllable or the second.

Production of Mandarin tones

Nagano-Madsen and Won (2017) conducted the study of production for the beginner group who participated in the perception experiment above. As for the production of L2 Mandarin tones, our prediction was the difficulty in manifesting pitch register since it is not the part of the Swedish pitch accent. It should be noted that speakers were consistent in the F0 manifestation of the target tone during their random repetition, which indicates that the tonal pattern in production is a stable 'interlanguage' at the time of investigation.

High-level tone (T1) was found to be the most successfully produced tone. Two out of nine students consistently produced the falling tone as rising tone. Figure 3 shows the example of the four tones that are differentiated only by using rising tones. As for the manifestation of T2 and T3, three different ways were observed. The first is to utilize both pitch register/glottalization and timing of the F0 rise like a native speaker. The second, which is most common among the Swedish learners, is to use only durational difference by excluding the F0 drop for T3. An example is shown in Figure 4 below. The third is to use only the difference in pitch register. Although glottalization gives a powerful cue in perception, its use is limited in L2 production. Furthermore, Swedish students have difficulty in producing T4 in which pitch falls deeply. Instead, the falling F0 often stops in the half way as can be seen in Figure 4.

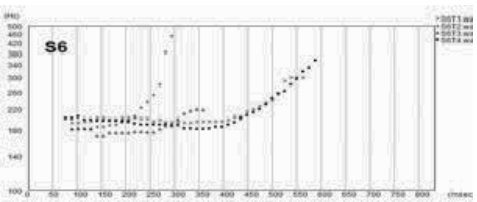

Figure 3. An example where all the four tones are differentiated by using only rising pitch.

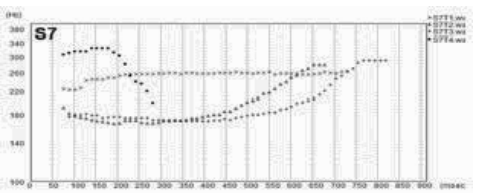

Figure 4. An example where T2 and T3 are differentiated only by duration.

Application to teaching Chinese tones to Swedish learners

In teaching the perception and production of Mandarin tones to the Swedish learners at the new beginner level, it is desirable to start with the difference between pitch rise and fall since many students found it difficult to discriminate the two. Students should be given a focused instruction for combination of tones that are difficult to them. In doing so, it would be desirable to demonstrate the F0 configurations to point out that the difference are mainly in the timing of F0 peak and valley as well as pitch registers. For example, the F0 configuration of T4+T2 pattern differs from T1+T2 pattern mainly in pitch register (Figure 5 below) while T2+T2 differs from T2+T3 in the timing of F0 fall and rise (Figure 6 below). The latter is similar to the manifestation of Swedish word accents.

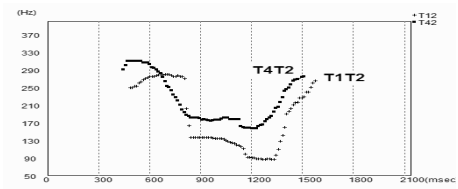

Figure 5. The two types of lexical pitch accents in Swedish. Male speaker.

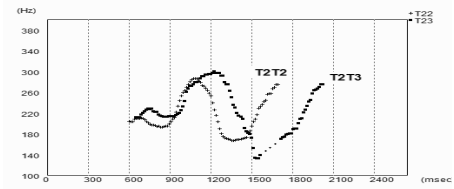

Figure 6. The two types of lexical pitch accents in Swedish. Male speaker.

Japanese

Word level prosody

Figure 7 below shows the F0 of two types of pitch accent in Japanese, the verb *noru* 'to ride' (flat/unaccented) and *nomu* 'to drink' (falling/accented). The falling accent can appear on any syllable in the word except for the last syllable, and the location is lexical, i.e. unpredictable.

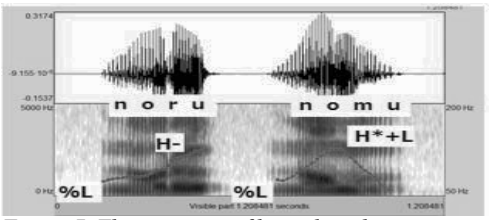

Figure 7. The two types of lexical pitch accents in Japanese.

Perception of Japanese pitch accent

Nagano-Madsen and Ayusawa (2002) conducted an accent identification test to examine how Swedish learners perceive accent in Japanese. Twenty-seven native Swedish students listened to 144 words varying in syllable structure, word length, accentual pattern, and intonation. Significant difference was found in discriminating the level accent type and other accent types and between the two groups of students who has studied in Japan or not. Particularly interesting is that Swedish learners do not perceive the difference between pitch rise and fall that appears utterance finally to indicate question vs. statement. In the test, the correct identification rate for statement and question was 43% and 47% respectively, with no statistical significance in the difference. Swedish learners have a tendency to perceive the Japanese accent as one mora delayed and to perceive the penultimate accent as flat (high level).

Production of Japanese pitch accent

L2 Japanese pitch accent distinction appears at the relatively early stage of acquisition. This may be due to the fact that Swedish also has a lexical pitch accent distinction (accent 1 and 2) although the exact phonetic realizations are very different from those in Japanese. Furthermore, there is a strong preference in placing an accent on the penultimate syllable (Nagano-Madsen 2014, 2015a). For the native Japanese pronunciation, there always is a sharp F0 fall directly after the accent while this sharp F0 fall is missing in the Swedish learners' utterances. Instead, there is a gradual F0 fall all the way to the end of a sentence. This phenomenon was reported earlier in Markham and Nagano-Madsen (1996) in which the participants were the Swedish students from southern Sweden.

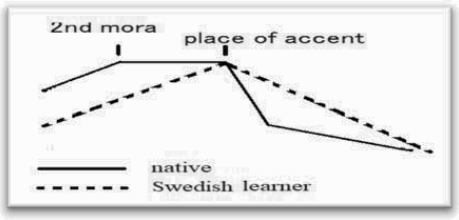

Figure 8. Comparison in producing the Japanese falling pitch accent, native vs. Swedish learners.

Phrasing and focus

A frequently used phrasing strategy by Swedish learners in reading Japanese text is discussed by Nagano-Madsen (2015b). The unique pattern found among the Swedish learners is that of 'upstep' instead of more common downstep in grouping words into a prosodic phrase. Compare Figure 9 and 10 below. In Figure 9, L1 intonation is shown for the sentence 'the old man went to the mountain to fetch twigs while the old woman went to the river to do washing'. Here, there are six words (AP=accentual phrases) that are grouped to two intermediate phrases (iPs). In Figure 10, an example of prosodic phrasing that is frequently found among the Swedish learner's speech but never in the native Japanese speech. It is exactly the same sentence shown in Figure 9. The phrasing strategy is the 'upstep' in which APs become successively higher in pitch, which is shown in the first half of Figure 10. Note this is exactly opposite of 'downstep' shown by a native Japanese speaker in Figure 9.

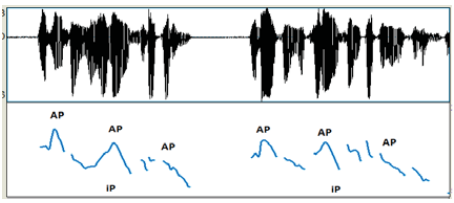

Figure 9. Phrasing by downstep, native Japanese speaker. 'Ojiisanwa jamae shibakarini, obaasanwa kawae sentakuni ikimashita (= the old man went to the mountain to fetch twigs while the old woman went to the river to do washing).

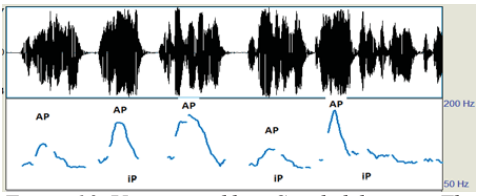

Figure 10. Upstep used by a Swedish learner. The same utterance as Figure 9.

Another dimension in which Swedish learners had difficulty in acquiring is the F0 features related to information/discourse structure (Nagano-Madsen, 2015a). Japanese is a topic-comment language where the topic is presented first with the topic particle wa. Then the new information follows. Figure 11 shows how the topic and focus are realized in a Japanese sentence 'the old man and woman gave the name MOMOTARO to the child'. The highest F0 peak is found for the word MOMOTARO whereas the F0 preceding this word is compressed. In the Swedish learners' utterances (Figure 12), the F0 relation is usually opposite, i.e. the topic phrase carries higher F0 than the focused phrase.

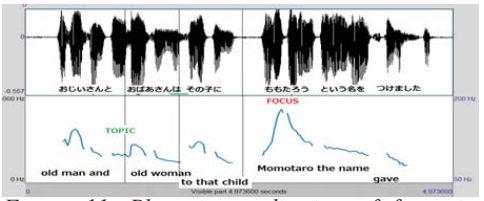

Figure 11. Phonetics realization of focus in Japanese. Native speaker.

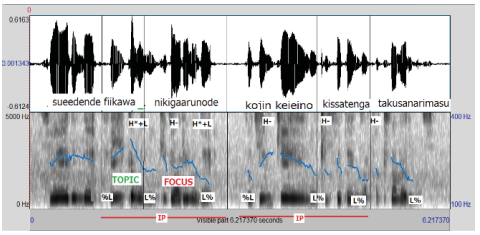

Figure 12. Phonetic realization of focus by a Swedish learner.

Application to teaching Japanese prosody to Swedish learners

Swedish learners did not succeed in differentiating a sentence final F0 rise vs. fall in the perception test. This may cause misunderstanding in communicating since Japanese questions are usually accompanied by a sentence final F0 rise. Furthermore, the

manifestation of focus needs to be instructed in relation to the topic marker particle 'wa', i.e. focus comes after this particle.

Conclusion

Both Japanese and Chinese have a contrast between a high-level pitch and a falling pitch. Although such a phonological contrast does not exist in Swedish, Swedish learners perform very well in discriminating the two both in perception and in production. In contrast, Swedish learners show a difficulty in discriminating a F0 rise from a F0 fall in perception. Swedish L2 prosody is characterized by the upward pitch movement, i.e. preference to use a F0 rise than to use a F0 fall in producing Chinese tones, preference of using upstep than to use downstep in prosodic phrasing of Japanese text, and there is a lack of a sharp F0 fall in their production of Chinese tones and Japanese pitch accent. In the recent pilot study comparing Swedish and Japanese prosody, it is hypothesized that Swedish learners acquire L2 intonational features before the respiratory control (Isei-Jaakkola, Nagano-Madsen, and Ochi, 2018).

References

Bruce G (1977). *Swedish word accents in sentence perspective.* Lund: Glerup.

Gårding E (1973). The Scandinavian word accents. *Working Papers* 8. Phonetics Laboratory, Lund University.

Isei-Jaakkola T, Nagano-Madsen Y, Ochi K (2018). Respiratory Control, Pauses, and Tonal Control in L1's and L2's Text Reading – A Pilot Study on Swedish and Japanese. *Proceedings of Speech Prosod.y*

Markham D, Nagano-Madsen Y (1996). Input modality effects in foreign accent. *Proceedings of ICSLP,* 3: 1473-1476. Philadelphia.

Nagano-Madsen Y, Bruce G (1998). Comparing pitch accent features in Swedish and Japanese. *Nordic Prosody*: 215-224.

Nagano-Madsen Y, Ayusawa T (2002). Analysis of the perceptual patterns of Japanese word accent by Swedish learners *Asia & Africa,* 2: 187-195. Department of Oriental and African Languages, University of Gothenburg.

Nagano-Madsen Y (2014). Acquisition of L2 Japanese intonation -data from Swedish learners (in Japanese). *Japanese Speech Communication,* 2: 1-27. Tokyo: Hitsuji Shobou.

Nagano-Madsen Y (2015a). Acquisition process of L2 Japanese intonation by Swedish learners - Interlanguage or prosodic transfer? *Proceedings of the18th ICPhS,* 2015:10-14. University of Glasgow.

Nagano-Madsen Y (2015b). Prosodic phrasing unique to the acquisition of L2 intonation- analysis of L2 Japanese intonation by L1 Swedish learners. *INTERSPEECH*, 100-104. Dresden, Germany.

Nagano-Madsen Y and Wang X (2016). Perception of L2 Mandarin Tones by Swedish Learners at Three Proficiency Levels. *Proceedings of the International Workshop on Language Teaching, Learning, and Technology*, 23-28. San Francisco, USA.

Nagano-Madsen Y & Wan X (2017). Perception and Production of L2 Mandarin tones by L1Swedish learners. *Proceedings of the European Signal Processing Conference*, 608-612. Kos, Greece.

Selinker L (1972). Interlanguage. *IRAL,* 10: 209-231.

Does understanding written language imply understanding spoken language for L2 users?

Monica Nyberg
Department of Swedish, Gothenburg University

Abstract

Results from two highstake L2 tests, the Dutch Staatsexamen NT2 and the Swedish Swedex, suggest that the level of reading comprehension does not necessarily agree with that of listening comprehension for L2 speakers. In this paper, an attempt is made to investigate whether the differences in reading and listening scores are not merely due to differences in vocabulary or text types being used in the tests. The results indicate that listening form a slightly more demanding activity for most of the participating students, something which is also confirmed by a complementary analysis carried out in the student group. One major reason is that interpretation of spoken language takes place in real time.

Introduction

Advanced users of an L2 may experience language as an obstacle while studying in their L2, and they seem to less often succeed in their studies than their native speaking peers (see for instance Deygers et al 2017, Zijlmans et al 2016 and Wijk-Andersson 2004), even though they have met the language entry requirements. Accordingly, there is reason to take a closer look at why this may be.

In Sweden, some L2 users prove their knowledge of Swedish by passing their L2 Swedish courses, and others by taking the Tisus test (*Test i svenska för universitets- och högskolestudier*). Different from many other countries' language entrance tests for university studies, Tisus does not include any listening comprehension test, but tests only speaking, writing and reading.

In my master's thesis (Nyberg to appear), my aim has been to explore whether a reading comprehension test can be said to cover for both receptive skills, or if a listening comprehension test would give valuable added information about a test taker's language proficiency level and language abilities necessary for university studies in Swedish.

The differences between spoken and written language may affect the ways an L2 user understand written and spoken texts and how large a vocabulary is needed (cf. Nation 2006 and Stæhr 2008). My hypothesis is therefore that the proficiency level for one of the two receptive skills does not necessarily agree with the other.,

Method

This study consists of two different comparisons between L2 users' reading and listening levels, accompanied by an analysis of L2 student essays on reading and listening. The first comparison is an analysis of test takers' results on reading and listening on two different highstake language tests at the B2 level: one Swedish exam (Swedex) and one Dutch language university entrance test (Staatsexamen NT2).

For the second comparison, I produced four reading and listening tests consisting of different texts given in written as well as in spoken form, and these were tested crosswise in a group of L2 speaking university students. The participating students should have a proficiency of Swedish at B2 level or higher. Texts from the Tisus reading comprehension test formed the base for testing the students on reading written texts as well as listening to recordings of them. Spoken texts, based on the Dutch Staatsexamen NT2 listening comprehension test, were used likewise, after translation and, for the listening version, recording. The reason behind this procedure is to, as far as it is possible, avoid that differences between test results for individual students depend on text type and vocabulary breadth.

Shortly after taking these tests, the participating students wrote a 300 words essay expressing their own ideas about their reading and listening needs and abilities, with main focus on the latter skill.

11 of the participating students participated in all of the tests and also handed in an essay afterwards.

Results

The first comparison between reading and listening levels showed that the two receptive skills does seem to match for most test takers involved, but not all. Out of 14 008 test takers in the Dutch Staatsexamen NT2, there were 1 315 failing their reading exam (but not their listening exam), and 1 221 failing on listening but not reading. This means that more than 18 % failed on the one receptive skill *or* the other. The equivalent number for the Swedex test was close to 16 %. 15 test takers out of 297 failed their reading comprehension exam (but not their listening) and 32 test takers failed on listening but not reading.

My experimental study similarly showed that the students' average results were clearly higher for the reading versions of the tests than for the listening versions (see Table 1). There were, however, many individual differences and due to the small number of participants it is not possible to draw any conclusions from this study.

Table 1. Average results on written text (reading and listening) and spoken text (reading and listening)

Reading	Listening	Reading	Listening
70,9%	61,8 %	73,1 %	59,8 %

From the 11 participating students' essays it is clear that they all are eager to understand spoken Swedish, at work/university as well as in their private life. However, they all shared experiences of having difficulties understanding spoken Swedish. Background noice, use of dialects, fast and unclear speech and difficult vocabulary are mentioned by several students as complicating factors. All students also share the feeling of not being able to interpret what they hear fast enough. Spoken language is transient and the listener must, automatically and in real time, interpret incoming messages at the same pace that the speaker speak (Buck 2001). All students but three expressed that reading was much easier than listening and those three students said that the level of perceived difficulty rather depended on the text type than the media. The perhaps most important difference between reading and listening for L2 students is that as a reader, they have some control: they can skim through some parts of the text and re-read others. While listening to a lecture, they need to keep up with the pace chosen by the lecturer. One of the students, however, reported that he made recordings of the lectures he attended and listened to them several times at home in order to understand what had been said. A large part of the group also claimed they had not practised any listening comprehension during their previous Swedish courses, or very little. They seem to have taken responsibility for approving their listening skills at home, mainly by watching Swedish television and listening to Swedish radio.

Discussion

Does understanding written language imply understanding spoken language? Yes, to a certain extent. Vocabulary and knowledge of syntax and grammar are needed for both receptive skills. There are, however, differences between written language and spoken languge. Written text may well be of a more complicated structure or perhaps contain more of word variation which would demand a larger vocabulary of the reader, but spoken language can be difficult due to unexpected pronounciation and speed of speech. While my study does not provide evidence that our Swedish university entrance test Tisus is in need of a listening comprehension test, it does give reason to further investigate the matter. How do the L2 students in Swedish manage? Are they ready, language wise, for not only reading course literature but also for listening to lectures? Some attention to listening comprehension would have beneficial consequences, if not to form a more valid and reliable language entrance test, then to produce washback effects on language teaching by encouraging some focus on the listening skill, which is otherwise easily forgotten in class.

References

Buck, G (2001). *Assessing Listening*. Cambridge University Press.

Deygers, B, Van den Branden, K & Van Gorp, K (2017). University entrance language tests: A matter of justice. *Language Testing*, 1-28.

Nation, P (2006). How large a vocabulary is needed for reading and listening? In: *Canadian Modern Language Review*, 63(1), 59-82.

Nyberg, M (to appear). Testa hörförståelse – ska det vara nödvändigt? Master's Thesis. Göteborgs universitet.

Stæhr, S (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign

language. *Studies in second language acquisition,* 31(4)-577-607.

Wijk-Andersson, E (1994). Rikstestresultat och studieframgång. Språk – utvärdering – test. Rapport från ASLA:s höstsymposium Karlstad 10–12 november 1994, 34:2, 163–173.

Zijlmans, L, Neijt, A & van Hout, R (2016). The role of second language in higher education: A case study of German students at a Dutch university. *Language learning in higher education*, 6(2), 473-493.

A Dual Complexity Gradient Theory of Speech Perception

Mikael Roll
Center for Languages and Literature, Lund University

Abstract

This article proposes a Dual Complexity Gradient theory of speech perception in the brain. The theory unifies a previously proposed phonetic-to-phonological complexity gradient along the ventral auditory processing stream in the temporal lobe with a recently suggested cortical structure complexity gradient. Findings supporting the theory are discussed as well as its predictions.

Introduction

To improve methods of language teaching and rehabilitation in a society with increasing linguistic diversity and a growing elderly population, we need fundamental information about the way native speakers process language. However, to date, the very fundamental process of how the brain integrates acoustic information into phonetic features, phonemes, syllables, and words is still not completely understood. We propose a theoretical framework, the Dual Complexity Gradient theory of speech perception, which can be used to generate new hypotheses about the relation between brain structure and phonetic processing. The Dual Complexity Gradient theory unifies a previously proposed phonetic-to-phonological complexity gradient along the ventral auditory perception stream with a recently suggested cortical structure gradient for phonetic-to-phonological processing. As preliminary evidence, it takes recent findings of correlation between native phonological proficiency and cortical thickness in different brain areas.

Auditory perception streams

Speech processing in the brain proceeds in two different streams (Saur, Kreher, Schnell, Kümmerer, Kellmeyer, Vry, Umarova, Musso, Glauche, Abel, Huber, Rijntjes, Hennig, & Weiller, 2008) most strongly represented in the left hemisphere (Peelle, 2012). Both streams begin in primary auditory cortex. The *dorsal stream* then extends to the superior temporal gyrus and through the parietal lobe via the superior longitudinal fasciculus to frontal cortex. The connection between auditory regions in temporal cortex and motor cortices in the frontal lobe is crucial for language learning, subvocal rehearsal (Buchsbaum, Olsen, Koch, & Berman, 2005), effortful processing, and predictive processes (Rauschecker & Scott, 2009; Roll, Söderström, Frid, Mannfolk, & Horne, 2017) possibly involving emulation (Grush, 2004). A parallel pathway through the arcuate fasciculus has been argued to be more involved in combinatorial syntactic parsing (Friederici, Chomsky, Berwick, Moro, & Bolhuis, 2017). The primary functions of the dorsal stream are sound localization and auditory sensorimotor integration (Rauschecker & Scott, 2009).

The *ventral stream* extends laterally and frontally from Heschl's gyrus to the anterior superior temporal gyrus and even the anterior superior temporal sulcus. From the anterior temporal lobe this stream continues to the ventral part of the frontal lobe through the extreme capsule. The basic function of the ventral stream is auditory object identification based on increasingly complex analysis of acoustic features (Rauschecker & Scott, 2009).

Spatial complexity gradient

A meta-analysis involving over 100 imaging experiments found a spatial phonetic complexity gradient along the ventral stream (DeWitt & Rauschecker, 2012). Thus, the further away one travels from primary auditory cortex along the ventral stream, the more complex the phonetic-to-phonological representations become, involving phonetic features, phonemes, syllables, words, and even short phrases. Primary auditory cortex has a rather detailed representation of the spectrotemporal characteristics of sounds. It is even tonotopically organized, meaning that different center frequencies map to different locations of the cortex (Humphries, Liebenthal, & Binder, 2010; Merzenich & Brugge, 1973).

Secondary auditory cortex (Fig. 1) has neurons that are sensitive to specific constellations of the spectrotemporal characteristics detected in primary cortex. These correspond to distinctions in manner of articulation—further organized into larger groups of sonorants and obstruents—and place of articulation. Secondary cortex also combines formants, giving rise to perception of vowel distinctions (Mesgarani, Cheung, Johnson, & Chang, 2014). Secondary auditory cortex further has normalized neural representations of tone and intonation patterns independent of absolute F0 level (Tang, Hamilton, & Chang, 2017). Tertiary auditory cortex, anteriorly located in the superior temporal gyrus, is sensitive to categorical auditory object perception in both humans (Patterson, Uppenkamp, Johnsrude, & Griffiths, 2002) and rhesus monkeys (Tsunada, Lee, & Cohen, 2011). For speech, this region seems sensitive to recognition of units at the word level. Slightly ventrally, in the frontal part of the superior temporal sulcus, even shorter phrases seem to be processed.

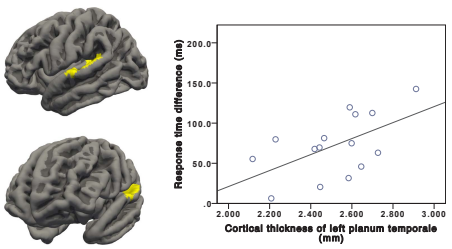

Figure 1. Cortical thickness of left secondary auditory cortex (planum temporale) in superior temporal gyrus correlates with proficiency in native language word accent processing (from Schremm, Novén, Horne, Söderström, Westen, & Roll, 2018).

Structural complexity gradient

The other part of the Dual Complexity Gradient theory is a *structural complexity gradient*, a generalization emanating from recent findings about the relation between brain structure and behavior. The structural gradient assumes that complex cognitive functions are aided by thicker cortex in brain areas modulating associative processing, whereas effective pruning and cortical myelination help lower-level cognitive functions, yielding instead an advantage of thinner cortex in primary processing areas (Novén, Schremm, Nilsson, Horne, & Roll, submitted).

Previously, gray-matter volume was the most commonly used structural brain metric. However, cortical volume indistinctly hides two measures that recent research suggests might be rather independently associated with cognitive abilities: cortical thickness and cortical surface area (Vuoksimaa, Panizzon, Chen, Fiecas, Eyler, Fennema-Notestine, Hagler, Franz, Jak, Lyons, Neale, Rinker, Thompson, Tsuang, Dale, & Kremen, 2016). Cortical surface area is strongly genetically defined (Vuoksimaa, Panizzon, Chen, Fiecas, Eyler, Fennema-Notestine, Hagler, Fischl, Franz, Jak, Lyons, Neale, Rinker, Thompson, Tsuang, Dale, & Kremen, 2015). Areas early in the auditory stream, primary and secondary auditory cortex, have greater surface area in the left hemisphere than in the right. Left primary auditory cortex is also thinner (Meyer, Liem, Hirsiger, Jäncke, & Hänggi, 2014). The reason is thought to be that a larger surface of well-organized cortical columns with more myelinated axons increases processing speed but reduces cortical thickness. This would be advantageous for the kind of rapid categorization involved in phonetic feature processing in left primary and possibly secondary auditory cortices (Long, Wan, Roberts, & Corfas, 2018; Warrier, Wong, Penhune, Zatorre, Parrish, Abrams, & Kraus, 2009). In line with this idea but outside the language domain, amusic individuals have been found to have thicker cortex than control persons in right primary auditory areas and inferior frontal gyrus (Hyde, Lerch, Zatorre, Griffiths, Evans, & Peretz, 2007). In the same vein, cortical thickness in right inferior frontal gyrus has been found to correlate negatively with pitch discrimination proficiency, which can be argued to be a low-level acoustic task (Novén et al., submitted).

Cortical thickness increases in response to mental training (Román, Lewis, Chen, Karama, Burgaleta, Martínez, Lepage, Jaeggi, Evans, Kremen, & Colom, 2016) and is negatively associated with aging (Thambisetty, Wan, Carass, An, Prince, & Resnick, 2010) and cognitive decline in degenerative diseases (Gerrits, van Loenhoud, van den Berg, Berendse, Foncke, Klein, Stoffers, van der Werf, & van den Heuvel, 2016). Although maturation in children in general implies cortical thinning (Porter, Collins, Muetzel, Lim, & Luciana, 2011), regions in higher-level areas in the left ventral speech processing stream rather increase in thickness during childhood (Sowell, Thompson, Leonard, Welcome, Kan, & Toga, 2004).

Secondary auditory cortex has shown seemingly paradoxical results for cortical thickness. Thus, a low-level auditory experiment found

increased electrophysiological response for *thinner* cortex (Liem, Zaehle, Burkhard, Jäncke, & Meyer, 2012). For more complex tone-suffix association, however, processing speed was seen to augment with *thicker* cortex (Schremm et al., 2018) (Fig. 1). Further, intense language training increases cortical thickness in the anterior portion of secondary auditory cortex (Mårtensson, Eriksson, Bodammer, Lindgren, Johansson, Nyberg, & Lövdén, 2012). In line with the proposed structural gradient, this likely indexes acquisition of a new phonology with novel combinations of phonetic features. The structural complexity gradient provides an explanation for the apparent paradox of cortical thickness advantage or disadvantage if secondary auditory cortex is an intermediate area between primary auditory cortex and higher-level cognition areas, as its myelination patterns seem to indicate (Glasser & Van Essen, 2011). Thus, as mentioned above, increased cortical myelination reduces cortical thickness and is related to faster processing of low-level features. However, higher-level processing requires increased association between information types, and is expected to be related to increased cortical thickness. Accordingly, primary cortices are highly myelinated whereas secondary cortices show an intermediate degree of myelination, and association areas, a low degree of myelination (Glasser, Coalson, Robinson, Hacker, Harwell, Yacoub, Ugurbil, Andersson, Beckmann, Jenkinson, Smith, & Van Essen, 2016).

Anterior to secondary auditory cortex, tertiary auditory cortex is thicker on the left side (Meyer et al., 2014). This is in accordance with accumulation of knowledge regarding high-level phonological patterns during language acquisition resulting in increased number of neurons, synapses, and/or glial cells (Zatorre, Fields, & Johansen-Berg, 2012). Thickness of tertiary auditory cortex is positively associated with speed of processing word accents in real words (Novén, Schremm, van Westen, Horne, & Roll, in preparation), possibly further indicating improved whole form storage in line with Schremm et al. (2018). When listeners were forced to use combinatorial processing to access tone-suffix associations in pseudowords, cortical thickness of Broca's area in the left inferior frontal gyrus rather correlated with speed of access (Fig. 2).

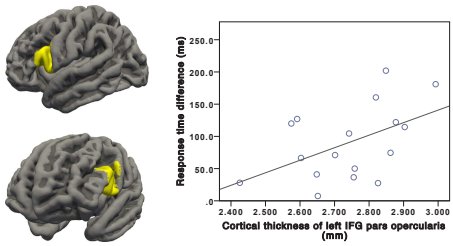

Figure 2. Cortical thickness of pars opercularis of left inferior frontal gyrus (Broca's area) with proficiency in native language word accent processing when combinatorial processing is forced by using pseudowords (from Schremm et al., 2018).

Native speaker proficiency

Previously brain structure correlates of phonetic or phonological proficiency had mostly been measured along the dorsal stream (Golestani, 2012). This is not strange since proficiency measures are mostly relevant for language learners. As mentioned above, the sensorimotor integration functions of the dorsal stream are crucial for language learning. However, recently, the speed of associating word accent tones with word endings in native speakers of Swedish begun to be measured (Roll, Söderström, & Horne, 2013; Roll, Söderström, Mannfolk, Shtyrov, Johansson, van Westen, & Horne, 2015; Söderström, Horne, Mannfolk, Westen, & Roll, 2017; Söderström, Horne, & Roll, 2017; Söderström, Roll, & Horne, 2012). Since this phonological association seems to be important for online prediction and facilitation in speech processing (Söderström, Horne, Frid, & Roll, 2016), dominating it can be seen as an indication of increased language proficiency, even in native speakers. This has made it possible to assess the relation between cortical thickness and a tentative measure of native phonological "proficiency" at different levels, giving rise to correlations along the ventral speech processing stream (Schremm et al., 2018).

Conclusions

This article has reviewed some works showing evidence for a previously suggested phonetic-to-phonological *spatial complexity gradient* in the ventral stream of auditory processing. This gradient proceeds in anterior direction form primary auditory cortex through secondary and tertiary cortices in the superior temporal gyrus and anterior superior temporal sulcus. The article has also taken up a recently suggested *structural*

complexity gradient, by which processing of low-level acoustic features is facilitated by a thinner and more myelinated cortex, where straightforward choices can rapidly be processed. Higher-level phonological processing is rather aided by a higher number of associations in a more complex network yielding thicker cortex. Further evidence is needed to corroborate the structural complexity gradient along the ventral speech perception stream, but unification of the two complexity gradients into one framework, a Dual Complexity Gradient theory, gives a number of testable predictions. Thus, proficiency at different levels of processing in one's native language should correlate with cortical thickness in different ways and in different brain areas. For lower-level processing and in primary areas, a negative correlation would be expected. This has been found to a certain degree in the right hemisphere for proficiency in non-speech pitch discrimination. Proficiency in higher-level processing would be thought to correlate positively with cortical thickness. This is what has been found for proficiency in word accent processing in secondary auditory cortex and, when involving forced combinatorial processing in pseudowords, in inferior frontal gyrus. Many points along the ventral pathway and levels of phonetic-to-phonological complexity need to be tested for the theory to be considered to be corroborated.

Acknowledgments

This work was supported by Knut and Alice Wallenberg Foundation (grant number 2014.0139) and Marcus and Amalia Wallenberg Foundation (grant number 2014.0039). I am grateful to Merle Horne, Mikael Novén, Andrea Schremm, Pelle Söderström and Sabine Gosselke Berthelsen for insightful discussion.

References

Buchsbaum BR, Olsen RK, Koch P, & Berman KF (2005). Human dorsal and ventral auditory streams subserve rehearsal-based and echoic processes during verbal working memory. *Neuron,* 48(4): 687-697.

DeWitt I, & Rauschecker JP (2012). Phoneme and word recognition in the auditory ventral stream. *Proceedings of the National Academy of Sciences,* 109(8): E505-E514.

Friederici AD, Chomsky N, Berwick RC, Moro A, & Bolhuis JJ (2017). Language, mind and brain. *Nature Human Behaviour,* 1(10): 713-722.

Gerrits NJHM, van Loenhoud AC, van den Berg SF, Berendse HW, Foncke EMJ, Klein M, Stoffers D, van der Werf YD, & van den Heuvel OA (2016). Cortical thickness, surface area and subcortical volume differentially contribute to cognitive heterogeneity in Parkinson's disease. *PLoS ONE,* 11(2): 1-14.

Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, Ugurbil K, Andersson J, Beckmann CF, Jenkinson M, Smith SM, & Van Essen DC (2016). A multi-modal parcellation of human cerebral cortex. *Nature,* 536: 171-181.

Glasser MF, & Van Essen DC (2011). Mapping human cortical areas in vivo based on myelin content as revealed by T1- and T2-weighted MRI. *Journal of Neuroscience,* 31(32): 11597-11616.

Golestani N (2012). Brain structural correlates of individual differences at low-to high-levels of the language processing hierarchy: A review of new approaches to imaging research. *International Journal of Bilingualism,* 18(1): 6-34.

Grush R (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences,* 27(3): 377-396.

Humphries C, Liebenthal E, & Binder JR (2010). Tonotopic organization of human auditory cortex. *Neuroimage,* 50(3): 1202-1211.

Hyde KL, Lerch JP, Zatorre RJ, Griffiths TD, Evans AC, & Peretz I (2007). Cortical Thickness in Congenital Amusia: When Less Is Better Than More. *Journal of Neuroscience,* 27(47): 13028.

Liem F, Zaehle T, Burkhard A, Jäncke L, & Meyer M (2012). Cortical thickness of supratemporal plane predicts auditory N1 amplitude. *Neuroreport,* 23: 1026-1030.

Long P, Wan G, Roberts MT, & Corfas G (2018). Myelin development, plasticity, and pathology in the auditory system. *Developmental Neurobiology,* 78(2): 80-92.

Merzenich MM, & Brugge JF (1973). Representation of the cochlear partition on the superior temporal plane of the macaque monkey. *Brain Research,* 50(2): 275-296.

Mesgarani N, Cheung C, Johnson K, & Chang EF (2014). Phonetic feature encoding in human superior temporal gyrus. *Science,* 343: 1006-1010.

Meyer M, Liem F, Hirsiger S, Jäncke L, & Hänggi J (2014). Cortical surface area and cortical thickness demonstrate differential structural asymmetry in auditory-related areas of the human cortex. *Cerebral Cortex,* 24(10): 2541-2552.

Mårtensson J, Eriksson J, Bodammer NC, Lindgren M, Johansson M, Nyberg L, & Lövdén M (2012). Growth of language-related brain areas after foreign language learning. *Neuroimage,* 63: 240-244.

Novén M, Schremm A, Nilsson M, Horne M, & Roll M (submitted). Cortical thickness of Broca's area and right homologue predict grammar learning aptitude and pitch discrimination proficiency.

Novén M, Schremm A, van Westen D, Horne M, & Roll M (in preparation). Cortical thickness of planum polare in native tone perception.

Patterson RD, Uppenkamp S, Johnsrude IS, & Griffiths TD (2002). The Processing of Temporal Pitch and Melody Information in Auditory Cortex. *Neuron,* 36(4): 767-776.

Peelle JE (2012). The hemispheric lateralization of speech processing depends on what "speech" is: a hierarchical perspective. *Frontiers in Human Neuroscience,* 6: 309.

Porter JN, Collins PF, Muetzel RL, Lim KO, & Luciana M (2011). Associations between cortical thickness and verbal fluency in childhood, adolescence,

and young adulthood. *Neuroimage,* 55(4): 1865-1877.

Rauschecker JP, & Scott SK (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature Neuroscience,* 12(6): 718-724.

Roll M, Söderström P, Frid J, Mannfolk P, & Horne M (2017). Forehearing words: Pre-activation of word endings at word onset. *Neuroscience Letters,* 658: 57-61.

Roll M, Söderström P, & Horne M (2013). Word-stem tones cue suffixes in the brain. *Brain Research,* 1520: 116-120.

Roll M, Söderström P, Mannfolk P, Shtyrov Y, Johansson M, van Westen D, & Horne M (2015). Word tones cueing morphosyntactic structure: neuroanatomical substrates and activation time course assessed by EEG-fMRI. *Brain and Language,* 150: 14-21.

Román FJ, Lewis LB, Chen C-H, Karama S, Burgaleta M, Martínez K, Lepage C, Jaeggi SM, Evans AC, Kremen WS, & Colom R (2016). Gray matter responsiveness to adaptive working memory training: a surface-based morphometry study. *Brain Structure and Function,* 221(9): 4369-4382.

Saur D, Kreher BW, Schnell S, Kümmerer D, Kellmeyer P, Vry M-S, Umarova R, Musso M, Glauche V, Abel S, Huber W, Rijntjes M, Hennig J, & Weiller C (2008). Ventral and dorsal pathways for language. *Proceedings of the National Academy of Sciences,* 105(46): 18035-18040.

Schremm A, Novén M, Horne M, Söderström P, Westen Dv, & Roll M (2018). Cortical thickness of planum temporale and pars opercularis in native language tone processing. *Brain and Language,* 176: 42-47.

Sowell ER, Thompson PM, Leonard CM, Welcome SE, Kan E, & Toga AW (2004). Longitudinal mapping of cortical thickness and brain growth in normal children. *Journal of Neuroscience,* 24(38): 8223.

Söderström P, Horne M, Frid J, & Roll M (2016). Pre-activation negativity (PrAN) in brain potentials to unfolding words. *Frontiers in Human Neuroscience,* 10: 1-11.

Söderström P, Horne M, Mannfolk P, Westen Dv, & Roll M (2017). Tone-grammar association within words: Concurrent ERP and fMRI show rapid neural pre-activation and involvement of left inferior frontal gyrus in pseudoword processing. *Brain and Language,* 174: 119-126.

Söderström P, Horne M, & Roll M (2017). Stem tones pre-activate suffixes in the brain. *Journal of Psycholinguistic Research,* 46: 271-280.

Söderström P, Roll M, & Horne M (2012). Processing morphologically conditioned word accents. *The Mental Lexicon,* 7(1): 77-89.

Tang C, Hamilton LS, & Chang EF (2017). Intonational speech prosody encoding in the human auditory cortex. *Science,* 357: 797–801.

Thambisetty M, Wan J, Carass A, An Y, Prince JL, & Resnick SM (2010). Longitudinal changes in cortical thickness associated with normal aging. *Neuroimage,* 52(4): 1215-1223.

Tsunada J, Lee JH, & Cohen YE (2011). Representation of speech categories in the primate auditory cortex. *Journal of Neurophysiology,* 105(6): 2634-2646.

Vuoksimaa E, Panizzon MS, Chen C-H, Fiecas M, Eyler LT, Fennema-Notestine C, Hagler DJ, Fischl B, Franz CE, Jak A, Lyons MJ, Neale MC, Rinker DA, Thompson WK, Tsuang MT, Dale AM, & Kremen WS (2015). The Genetic Association Between Neocortical Volume and General Cognitive Ability Is Driven by Global Surface Area Rather Than Thickness. *Cerebral Cortex,* 25(8): 2127-2137.

Vuoksimaa E, Panizzon MS, Chen C-H, Fiecas M, Eyler LT, Fennema-Notestine C, Hagler DJ, Franz CE, Jak AJ, Lyons MJ, Neale MC, Rinker DA, Thompson WK, Tsuang MT, Dale AM, & Kremen WS (2016). Is bigger always better? The importance of cortical configuration with respect to cognitive ability. *Neuroimage,* 129: 356-366.

Warrier C, Wong P, Penhune V, Zatorre R, Parrish T, Abrams D, & Kraus N (2009). Relating structure to function: Heschl's Gyrus and acoustic processing. *Journal of Neuroscience,* 29(1): 61-69.

Zatorre RJ, Fields RD, & Johansen-Berg H (2012). Plasticity in gray and white: neuroimaging changes in brain structure during learning. *Nature Neuroscience,* 15: 528-536.

Pronunciation of foreign names in public service radio: How can phonetic transcriptions be of help?

Michaël Stenberg
Centre for Languages and Literature, Lund University

Abstract

Radio stations have different policies regarding pronunciation of foreign names. Over the years, the BBC (British Broadcasting Corporation) and Sveriges Radio, its Swedish equivalent, have taken great care with pronunciation, particularly of foreign names. However, broadcasters cannot be expected to possess the skills or knowledge needed to do their job without consulting experts, books, or other sources of information. Among those are online guides that may contain sound recordings as well as phonetic transcriptions. The present paper discusses to what extent phonetic transcriptions can be helpful for public service staff and what kind of transcriptions would be most advantageous.

Background

The British Broadcasting Corporation (BBC), founded in 1922 as British Broadcasting Company Ltd., and Sveriges Radio, its Swedish equivalent, commencing in 1925 as Radiotjänst, are two Public Service radio companies that have been forerunners with respect to pronunciation of names, notably foreign ones. For a long time, the BBC domestic services have endeavoured to pronounce geographical as well as personal names in a way close to that of the original languages.

Why bother about pronunciation?

There are several reasons for bothering about pronunciation in radio transmissions. One is, of course, intelligibility. Another one – not less important – is credibility: persons who are talking about politics, music, literature, film, etc. and not knowing how to pronounce the names of central protagonists, might convey to listeners the impression that they are not well-acquainted with the field, but merely have been reading about it of late. A further reason can be to safeguard the prestige of the broadcasting company as a serious and reliable medium.

In The Guardian (11 March 2014) Steven Poole writes: 'If I say "SKED-ule" and you say "SHED-ule", will any farcical misunderstandings or tragic loss of life ensue? … but that was never the point. Like so many linguistic arguments, the power-struggles over correct pronunciation are most often proxies for issues of snobbery and class. The completely unpredictable pronunciations of many proper names in English, for example, act as a kind of secret code for the elect.'

IPA versus other transcription systems

One would imagine that IPA were the preferable transcription system for providing broadcasters with information on pronunciation, now that it is being increasingly used – certainly in slightly varying versions – in Wikipedia (whose Swedish language edition with its more than 3,700,000 entries is the second biggest after the English one). However, at least in Sweden, where the teaching of foreign languages has suffered a decline since the 20th century, it seems less viable than before. Teachers in Swedish compulsory and upper secondary schools apparently do not take issues of pronunciation seriously; most probably, they do not let their pupils acquaint themselves with the IPA. No wonder that a great deal – not to say a majority – of Swedish speakers in Public Service radio pronounce the English word *national* as [ˈnætʃənəl].

BBC and Sveriges Radio – a comparison

Early in the history of the BBC, a dedicated pronunciation unit was created. Housed in Broadcasting House, Portland Place, London W1, in its heyday it comprised a good dozen people, according to its director Graham Pointon, interviewed by the present author in 1992. Since then, the number of employees has diminished. According to an interview with Mr. Pointon in the New York Times, made in the same year, the start of the pronunciation research unit dates from 1924, when the Welsh phonetician Arthur Lloyd James (1884–1943) was called upon to lecture for the announcers, who learnt so much from him that they asked for a full-time pronunciation adviser.

Over half a century of continuous research resulted in a card index made up of about 250,000 items. Based on this and on further research, Mr. Pointon, at the service of the BBC staff and the general public, has edited a pronouncing dictionary of British names (1993). In this, for every entry pronunciation is given in two transcription systems: IPA and a phrase-book-like system featuring only letters of the English alphabet. Referring to foreign names or words, the latter system brings about a slight anglisizing. The same principle is applied by two of Pointon's successors in their Oxford BBC Guide to pronunciation (Olausson & Sangster, 2006).

In Stockholm, the corresponding activitities at Sveriges Radio – until 1957 called Radiotjänst – originated as a card index established in the 1950's and '60's by a few engaged announcers, says Stefan Lundin in an interview made by the present author in 2013. Lundin himself was employed 1992–2015 as a linguistic adviser at Sweden's three joint Public Service companies Sveriges Radio, Sveriges Television and UR (educational radio & TV). The card index was created to support announcers of music and culture programmes, where names of authors and actors figured prominently. About 3,000 recommendations are included in the index.

The tradition since the beginning was to use a simplified transcription system departing from the Swedish alphabet, with certain signs added. To use a simple system seemed to be the right way to go dealing with radio and television co-workers.

Since 1981, a publication named Språkbrevet (The Language letter) is being continuously distributed, initially as a leaflet, from 2014 on in pdf format – at no charge for the interested general public. Pronunciation advice is an important part of its contents. A recent innovation is Dixi, a database accessible exclusively on the joint intranet of the Public Service companies. Dixi uses the same transcription system as Språkbrevet, i.e. one based on the letters of the Swedish alfabet and their normal pronunciation, but with addition of certain IPA symbols (ə ɑ θ ʃ ʒ x), which are explained in direct connexion to each occurrence.

Here are some examples:

David Guetta /dav**i**dd gett**a**/ fransk musikproducent och dj

David Rautio, /d**ɑ:**vid r**a͡o**tiå/ inget s-ljud efter *t*:et svensk ishockeymålvakt

Helene Schjerfbeck /hele:n ʃ**ä**rvbäkk/ finländsk konstnär, obs! Inget f-ljud i *Schjerfbeck*

Majdanek /majd**a**nekk/, tryck på andra stavelsen nazistiskt förintelseläger under andra världskriget

Tonio Borg /tå:niå bårrt͡tj/ EU-kommissionär från Malta

cache /k**a**ʃ, rimmar på krasch, [...] tillfällig lagringsplats i datorn
plural: *cachar* cacheminne

Li Keqiang /li kə t͡tjang/, /ə/ som i eng. *the* Kinas premiärminister

Internally, the announcers' corps, where Lundin was working, made use of IPA, specifically the Duden notation, as in Duden Aussprache- wörterbuch.

One of the major things that Lundin did before retiring was reading aloud and simultaneously recording all important items contained in Dixi, thus giving the staff easy access to the recordings via the intranet. The recorded version constitutes the recommendation.

Online sources of pronunciation information

On the internet there are sites in abundance that pretend to give information or recommendations on pronunciation issues, both sound recordings and various kinds of phonetic transcriptions. Many are managed by sheer amateurs, but among the more reliable – though using phrasebook style transcriptions – are *VOA Pronunciation Guide* and *You say it how?*
(The acronym VOA stands for Voice of America, a US Government Radio station, founded in 1942, transmitting to other countries in many languages, a propaganda medium, though less aggressive than Radio Free Europe/Radio Liberty.)

Conclusion

Phonetic transcriptions could certainly be of help for broadcasters, provided they are understandable and not too advanced. One might think that transcriptions would be superseded by sound recordings, but, as a matter of fact, they might often serve as a complement to the recordings. Some individuals are particularly skilful at imitating recordings, even in languages unknown to them; others need visual support. If IPA were taught on a large scale in schools, it could be substituted for the many inferior systems currently in use.

References

Printed resources

BBC Pronouncing dictionary of British names, 2nd edn (1983). Pointon, G E (ed) Oxford: Oxford University Press.

Brink, L et al. (1991). *Den Store Danske Udtaleordbog.* København: Munksgaard.

Duden, Aussprachewörterbuch, 6th edn (2005). Mannheim: Dudenverlag

Garlén, Claes (2003). *Svenska språknämndens uttalsordbok.* Stockholm: Svenska språknämnden: Norstedts ordbok.

Olausson L & Sangster C (2006). *Oxford BBC Guide to pronunciation: the essential handbook of the spoken word.* Oxford: Oxford University Press

Siebs, Th (1931). *Rundfunkaussprache.* Berlin: Reichs-Rundfunk-Gesellschaft (als Handschrift gedruckt).

Wells, J C (2008). *Longman pronunciation dictionary, 3rd ed.* Harlow, England: Pearson Education Ltd.

Internet resources

VOA Pronunciation Guide
https://www.insidevoa.com/a/pronunciation-guide-143860776/178765.html

You say it how?: LBPH Pronunciation Guide to Names of Public Figures. Jefferson, Missouri, USA: Missouri Secretary of State's web site
https://www.sos.mo.gov/wolfner/sayhow/a

Durational properties of word-initial consonants – an acoustic and articulatory study of intra-syllabic relations in a pitch-accent language

Malin Svensson Lundmark
Centre for Languages and Literature, Lund University

Abstract

The present study investigastes systematic duration variation of word-initial consonants as a function of the binary tonal Swedish word accent distinction (high tone and low tone, respectively). Results reveal acoustic durational differences, as well as anticipatory articulatory constriction differences, between the two word accents. Moreover, the study also exposes systematic durational discrepancy between the anticipatory movements and their phonemic targets.

Phonetic research has shown that subtle speech production variation between languages results in systematic differences in phonemic targets (Foulkes et al., 2013). Phonemic systematic variation can also be observed within the same language and within the same dialect. For instance, segment duration has been shown to not only be affected by well-known factors such as speech rate, word frequency, prominence, and boundary signaling, among others (Fletcher, 2013), but also systematically vary as an effect of intra-syllabic relations. For example, in an acoustic study on Italian, a systematic variation was found in word-initial consonants, which proved to differ in duration depending on whether the adjacent vowel was followed by a singleton or a geminate (Turco & Braun, 2016). Yet, not only phonotactics have a systematic effect on intra-syllabic segmental duration. In the pitch-accent language Swedish, intra-syllabic segment duration differences have been found between the binary distinctive word accents, in which either a falling tone or a rising tone is associated with the primary stress of a word. Previous studies on the word accents have revealed acoustic durational differences in the stressed vowel, as well as in the post-vocalic consonant, but thus far not in the word-initial consonant (Elert, 1964; Svensson Lundmark et al., 2017).

The present study hypothesizes an effect of the word accents on word-initial consonants based on recent findings from an articulatory study with 19 South Swedish speakers on the coarticulation of the consonant-vowel (CV) sequence /ma/ (Svensson Lundmark et al., n.d.). The study revealed that the anticipatory movements of the articulators (tongue body and lips), reaching for the phonemic targets of the CV sequence, were sensitive to the suprasegmental feature of the word accents. Thus, in the present study, apart from presenting acoustic duration results on the same data on the word-initial consonant /m/, a second aim is to compare the acoustic duration measures with the articulatory constriction measures, and find systematic correlation between the anticipatory articulatory movements and the duration of the phonemic targets.

Mixed effects regression models have been carried out on the word-initial consonant /m/ of target words with a /maCV(C)/-sequence, as spoken by the 19 speakers. Results reveal a systematic duration variation of the word-initial consonant between the two word accent conditions. A significant effect of the word accents is found in both the acoustic results (Figure 1), and in the data on the anticipatory articulatory measurements (Figure 2). Results on normalized acoustic duration further display possible effects by word frequency and higher-level prominence (Figure 3). Moreover, there is a systematic discrepancy concerning the different word endings between the anticipatory articulatory movement leading to the phonemic target and the acoustic duration of the word-initial consonant.

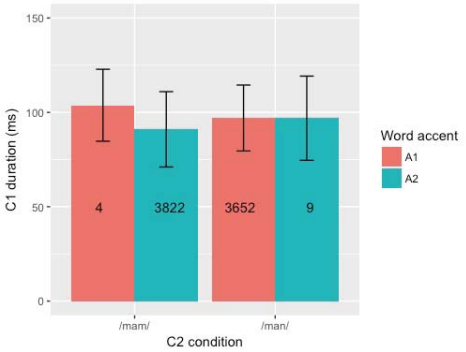

Figure 1. Acoustic duration of word-initial consonant [m] (C1), and the effect of word accent: A1 or A2, in target words where C2 is either a [m:] (/mammut/, /mamma/) or a [n:] (/mannen/, /manna/). Within each bar: word frequency according to the PAROLE corpus[1].

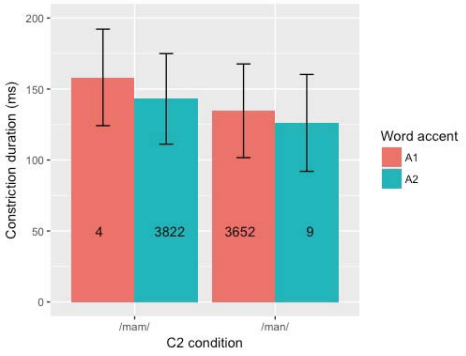

Figure 2. Anticipatory articulatory constriction duration (the bilabial closure) of the word-initial consonant [m] (C1), and the effect of word accent: A1 or A2, in target words where C2 is either a [m:] (/mammut/, /mamma/) or a [n:] (/mannen/, /manna/). Within each bar: word frequency according to the PAROLE corpus.

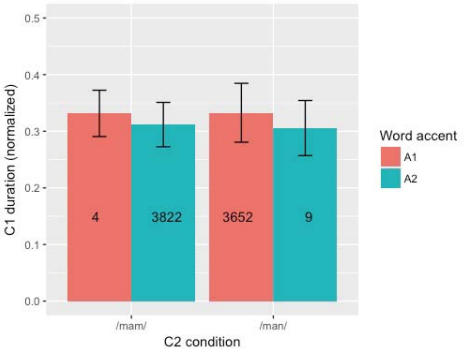

Figure 3. Normalized duration of the word-initial consonant [m] (C1), and the effect of word accent: A1 or A2, in target words where C2 is either a [m:] (/mammut/, /mamma/) or a [n:] (/mannen/, /manna/). Within each bar: word frequency according to the PAROLE corpus. The normalized duration is defined as a value from 0 to 1, where 1 equals the duration of the /maC:/-sequence.

References

Elert C-C (1964) *Phonologic Studies of Quantity in Swedish. Based on Material from Stockholm Speakers*. Sweden: Almqvist & Wiksell.

Fletcher J (2013) The Prosody of Speech: Timing and Rhythm. In: W J Hardcastle, J Laver, F E Gibbon, eds, *The Handbook of Phonetic Sciences: Second Edition.* UK: Wiley-Blackwell, 523–602.

Foulkes P, Scobbie J M, Watt D (2013) Sociophonetics. In: W J Hardcastle, J Laver, F E Gibbon, eds, *The Handbook of Phonetic Sciences: Second Edition*. UK: Wiley-Blackwell, 703–754.

Svensson Lundmark M, Frid J, Ambrazaitis G, Schötz S (n.d.) Word-initial CV coarticulation in a pitch-accent language. Manuscript submitted for publication.

Svensson Lundmark M, Ambrazaitis G, Ewald O (2017) Exploring multidimensionality: Acoustic and articulatory correlates of Swedish word accents. In: *Proc. of Interspeech 2017,* Stockholm, 3236–3240.

Turco G, Braun B (2016) An acoustic study on non-local anticipatory effects of Italian length contrast. *J. Acoust. Soc. Am.* 140: 2247–2256

[1] The Swedish PAROLE corpus. http://spraakdata.gu.se/parole/lexikon/swedish.parole.lexikon.html

Acoustic Results of Pronunciation Training

Bosse Thorén and Hyeseung Jeong
Department of Social and Behavioral Studies, University West

Abstract

The study examaines changes in a native Malaysian's pronunciation of word stress, vowel length and consonant clusters before and after she was tranined in English. Jeong & Thorén (under review) showed that this training made the speaker's pronunciation more intelligible to native Swedish listeners. We look at the changes on an acoustic level by measuring vowel duration, word duration, f0 patterns and realization of consonant clusters. The result shows that, after training, her pronunciation improved in all the three aspects. The improvement was audible as well as measurable with acoustical changes in absolute and relative vowel duration and realization of consonant clusters. Word stress patterns were also improved but with less clear-cut acoustic correlates.

Background

English is used as a Lingua Franca over the whole world and its non-native users now outnumber its native users (Jenkins 2015). Different varieties of English exhibit a great variety of different pronunciations.

Lately, a number of researchers (e.g. Jenkins, 2000, 2002, Derwing & Munro 2015) have promoted intelligibility between different varieties rather than aiming for either British Received Pronunciation (RP) or General American (GA) accent in educational contexts. Jenkins (2002) also proposed a number of phonetic features as Lingua Franca Phonetic Core for English, i.e. phonetic features that are crucial for intelligibility. Vowel length and consonant clusters are two of the phonetic features proposed by Jenkins. Word stress patterns are not included in Jenkin's proposal but were found important for intelligibility by Field (2005), Kashiwagi, Snyder, & Craig (2006) and Saito and Shintani (2016).

Based on the intelligibility principle, Jeong, Thorén and Othman's (2017) study examined the mutual intelligibility of Malaysian English and Swedish English. They found that native Malaysian listeners understood English spoken with Swedish accent better than native Swedish listeners understood English spoken with Malaysian accent. In order to test some of the suggested phonetic core features' impact on intelligibility, they designed 15 sentences, of which five aimed at testing the impact of word stress patterns, five to test long vowels and five to test consonant clusters. The sentences were assumed to be obviously true or false when understood. Malaysian and Swedish listeners responded by orally answering 'true', 'false' or 'I don't know' when hearing the sentences. Jeong, Thorén and Othman (under review) then let the same Malaysian speaker practice the three phonetic features in the 15 sentences mentioned above. The comprehension improved substantially when a new group of 21 Swedish listeners responded to the second version of the same 15 sentences.

The present study looks at the acoustic changes in the three phonetic features after the Malaysian speaker in the two aforementioned studies altered her pronunciation in a short training program. It aims find out what acoustical changes actually took place between the recordings after the training.

Typical properties of Malaysian English

From our own experiences on site (Kuala Lumpur) and the literature (Baskaran, 2004, 2008; Mesthrie and Bhatt, 2008; Yong, 2001), we conclude that Malay accented English has some deviations from standard English, although details vary among intra-national varieties e.g. Malay-Malay, Chinese-Malay and Indian-Malay. Some typical features shared by most native Malay speakers are omitting word final obstruents, reducing three-consonant clusters to two-consonant clusters, reducing two-consonant clusters into single consonants, and even omitting singleton final obstruents. The speech rhythm is more syllable-timed than standard English, meaning that unstressed syllables are not reduced in prominence or vowel quality and stressed syllables are seldom pronounced with longer

duration than unstressed ones. In addition to this, there is a general tendency to stress word final syllables in multi- and di-syllabic words.

Acoustic correlates of word stress, vowel length and consonant clusters in English

Fry (1955, 1958) found that duration was a more reliable cue to stress than intensity and that f0 was the most reliable cue. More recent studies, e.g. Morton & Jassem (1965), Eriksson & Heldner (2015) have mostly confirmed Fry's findings and also added the parameter of spectral emphasis as a major acoustic cue to lexical stress in English.

Correlates and measurements of vowel length and consonant clusters are expected to be more straightforward. Vowel length can be measured as vowel duration relative to word or utterance duration and realization of consonants can be checked by a combination of listening and examination of spectrograms.

Research questions

- Are there audible differences in all the target words between the two recordings?
- Will our chosen measurements correspond to all audible differences?
- Will temporal or tonal correlates dominate as acoustic realizations of intended and audible changes in word stress patterns?

Method

Speech material

The sentences used to test are listed below. The words aiming to test the respective phonetic features are underlined.

Sentences testing stress patterns

Vegetarians like to eat sausage salad.
The smallest animal in Africa is the elephant.
Military service is for women only.
A semester is a period in schools or universities.
A trumpet is a musical instrument.

Sentences testing consonant clusters

Kids wear glasses to walk fast.
Most birds make a nest to lay their eggs.
Ducks often swim in lakes and ponds.
Lots of textbooks describe facts.
(Boxers must use only their fists to strike each other.)

Sentences testing long vowels

We can feel with our feet when the floor is warm.
Birds can read from birth.
You often see leaves on trees.
Car seats must be made of steel.
Nobody wants peace on earth.

The sentences aiming at testing the impact of word stress patterns contained 2-4 words with 2-5 syllables each, giving opportunities for variation in stress placement. Sentences aimed at testing vowel length contained 2-4 monosyllabic words containing long/tense vowel in standard varieties of English and finally the sentences aiming at testing consonant clusters contained 2-4 words, each with 2-4 consecutive consonants. One of those sentences ('Boxers…') was omitted due to many listeners' problem with understanding its factual content.

Recordings

Recordings were made in a music studio at a university in Malaysia with professional equipment. Sample frequency was 48 kHz, the recordings were saved as Microsoft Wav files and edited and analysed in Praat (Boersma and Weenink, 2014).

The speaker

The speaker was a 21 year old female student at an English teacher education programme at a university in Malaysia. Her first language is Malay and English is her second. She speaks English with a moderate Malay accent.

Training

After the first recording, the speaker was trained to realize the three phonetic features more clearly, during an hour of tutoring by the first author and then practicing on her own for a week until the second recording. The recording session was preceded by a repetition of the tutoring.

The training focused solely on improving the specific features and only in the words that were intended to test those features. Everything else was left to her own intuitive pronunciation.

The recording entailed more than one version of the 15 sentences and we used the sentences that sounded most successful according to the intention and the instructions.

Measurements

We measured absolute and relative durations of vowels. The relative measurement is vowel duration divided by word duration. Checking realization of consonants in consonant clusters was done by a combination of listening and examining spectrograms. As improvement of consonant clusters we counted stops that changed from glottal to dental and also stops that changed from mere occlusion to occlusion plus burst. There were also word and syllable final obstruents that were totally unrealized – as far as we could detect – in the first version and clearly realized in the second version.

Measuring temporal word stress realization was performed by means of vowel duration or duration of vowel + voiced sonorant in the intended stressed syllable. Absolute durations were divided by word durations. Tonal realization was measured as follows: F0 gestures were identified and measured in vowel or sequence vowel + voiced sonorant belonging to the same syllable (*only, trumpet, instrument*). Direction of tonal change was registered as falling, rising or flat. Acoustical properties were compared to the authors' subjective judgement of whether the stress pattern was changed or not between the two recordings.

Methodological considerations

Our first plan was to measure duration of intended stressed syllables as well as f0 change in the same syllables in terms of pith range and pictch slope defined as semitones and semitones per second respectively. The picture however, was complicated by the fact that nearly all intended stressed syllables have longer absolute duration in recording 2 compared to recording 1. This means that a longer syllable and a larger pitch range could anyhow result in a less steep slope. Therefore we decided to identify tonal gestures by studying the visual pitch patterns in Praat (Boersma & Weenink 2014).

Result

Audible changes

The listeners in this sub-study are the present authors; one native Swedish and one native Korean. We however agreed nearly totally in all three categories. The cases we did not agree completely is actually only one case in the sentences aimed at testing word stress patterns. In the sentences testing vowel length, we heard increased vowel length in all 15 target words. In the sentences testing consonant clusters we heard improvements in 7 out of 13 target words. In the sentences aimed at testing word stress improvement we agreed that 13 out of 16 target words had improved stress patterns.

Realization of word stress

Out of a total of 16 target words, 13 words in the second recording were heard as having a more distinct correct stress pattern by both present authors. We both agreed that the words 'military' and 'sausage' had a clear and correct stress pattern in both versions with no audible difference. The word 'trumpet' was heard as improved by the first author but not by the second author. Three words; 'only', 'instrument' and 'salad' were judged as having clearly incorrect stress patterns in the first recording, stressing the last instead of the first syllable. The word 'vegetarians' was perceived as having stress on the first syllable in the first recording and on the third syllable in the second recording, Other words had generally unclear stress patterns in the first recording.

Tonal changes could be traced in all but one of the words in which we perceived an improved stress pattern and in four out of those, the tonal change was accompanied by an increase of duration. We also see from our data that sometimes a substantial change has taken place in the intended stressed syllable between first and second recording and sometimes we see more of changed internal relations within the word, e.g. adjacent syllables have become shorter or have decreased tonal prominence while the properties of the target syllable are more or less intact. This can be seen in the word 'semester' in figure 1.

Table 1 shows the three categories that emerged from the measurements: i) tonal change, ii) both temporal and tonal change and iii) the third category where we heard an improved stress pattern but did not find any change in our chosen acoustic measurements.

Figure 1 below is a crude modelling of f0 patterns for the target words in the first and second recording and shows a sample of different patterns in the first and second recording respectively.

Table 1. Groups of words with different or no observable correlates of stress.

Tonal change	Tonal + temporal change	No detected acoustic change
service	only	smallest
women	animal	
Africa	musical	
elephant	instrument	
vegetarians		
salad		
semester		
universities		

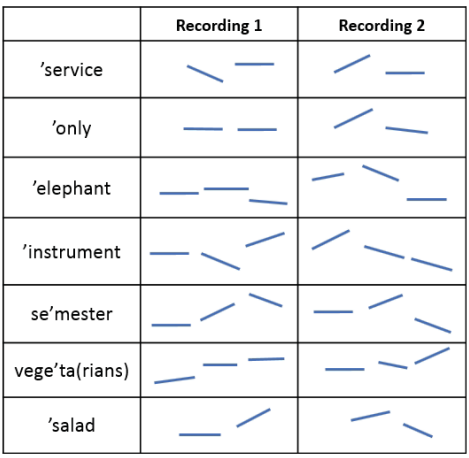

Figure 1. Samples of tonal patterns for words that were perceived as having improved stress pattern from the first to the second recording.

The bars in figure 1 illustrate the tonal gesture in the relevant syllables. In the case of 'vegetarians', the figure shows only the three first syllables including the intended stressed one. Subsequent syllables were not possible to parse. Different lengths of the bars is due to inprecise drawing and has no prosodic significance. Our results comply with those of Fry (1958) insofar as what concerns the direction and amplitude of the f0 change. Going from flat or falling to rising f0 seems to be a stronger perceptual cue to syllable prominence than the steepness of rises or falls.

Realization of consonant clusters

8 out of 13 words had addition of segment(s) or improved clearness of one segment. Examples of improved clearness was higher amplitude burst in /k/ in 'lakes' and going från glottal stop for /d/ in 'kids' to dental occlusion and burst. In the word 'textbooks', there were two added consonants: [ˈtæksˌbuk - ˈtækstˌbuks].

Realization of vowel length

13 out of 15 words had longer absolute and relative vowel duration in the second recording. The relative vowel duration was on average increased by 25% and the absolute duration was increased by some 160%.

Discussion and conclusion

In the vast majority of the target words, there were both audible and measurable changes leading to improved clearness. First of all we can state, in the case of vowel length and consonant clusters, that we always found a physical acoustic change when we heard one. The relation between perception and physical properties is rather straightforward. We acknowledge, however, that the speaker's effort to lengthen vowel sounds and clearly producing more consonants than she was used to, reduced her overall speaking rate, which in turn probably contributed to increased intelligibility.

Acoustic correlates to word stress is more complex and since we heard the word 'smallest' as having improved stress pattern and did not find any change in our acoustic observations, we have to suspect that there is at least one more acoustic variable involved. Since spectral emphasis was found by Eriksson & Heldner (2015) to be an influential correlate of stress, we believe that this could very well be the acoustic measure that our study lacks.

Concerning the third research question, we can clearly state that tonal change played a more significant role than temporal change. Thus our result complies with both Fry (1958) and Eriksson & Heldner (2015).

We hope that more knowledge in the areas of priority among phonologic and phonetic features as well as acoustic correlates to relevant phonologic properties can be beneficial in second language instruction, both human to human and computer to human. We need to know which phonetic features are more or less important for intelligibility and, especially in digital interactional systems, we need knowledge of the physical and measurable acoustics behind our intuitive linguistic perception.

References

Baskaran, L (2004). Malaysian English: Phonology. In B Kortmann & E W Schneider (Eds.), *A handbook of varieties of English*. Berlin, Germany: Mouton de Gruyter. 1034-1046

Baskaran, L (2008). Malaysian English: Phonology. In R Mesthrie (Ed.), *Varieties of English: Africa, South and Southeast Asia.* Berilin, Germany: Mouton de Gruter. 278-291.

Boersma, P, & Weenink, D (2014). Praat: Doing phonetics by computer. Version 5.4.04. Retrieved from http://www.praat.org/

Derwing, T, & Munro, M (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research* Amsterdam: John Benjamins.

Eriksson, A & Heldner, M (2015). The Acoustics of Word Stress in English as a Function of Stress Level and Speaking Style. *16th Annual Conference of the International Speech Communication Association (INTERSPEECH 2015)*, Dresden, Germany, September 6-10, 2015

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly,* 399-423

Fry D B (1958). Experiments in the perception of stress. *Language and Speech.* 126-152.

Jenkins, J (2000). *The phonology of English as an international language* Oxford, England: Oxford University Press.

Jenkins, J (2002). A sociolinguistically based, empirically researched pronunciation syllabus for english as an international language. *Applied Linguistics,* 83-103.

Jenkins, J (2015). *World Englishes: A Resource Book for Students.* London, England: Routledge.

Jeong H Thorén B and Othman J (2017). Mutual intelligibility of Malay- and Swedish- accented English: an experimental study. *Indonesian Journal of Applied Linguistics*, 43–57.

Jeong H Thorén B and Othman J (under review). Effect of altering three phonetic features on intelligibility of English as a lingua franca: A Malaysian speaker and Swedish listeners.

Kashiwagi, A, Snyder, M & Craig, J (2006). Suprasegmentals vs. segmentals: NNS phonological errors leading to actual miscommunication. *JACET Bulletin, 43*, 43-57.

Mesthrie, R, & Bhatt, R M. (2008). *World Englishes: The study of new linguistic varieties* Cambridge, England: Cambridge University Press.

Morton, J and Jassem W. (1965) Acoustic Correlates of stress. *Language and Speech.* 159-181.

Saito, K & Shintani, N (2016). Do Native Speakers of North American and Singapore English differentially perceive comprehensibility in second language speech? *TESOL Quarterly*. 421-446.

Yong, J Y (2001). Malay/Indonesian speakers. In M Swan & B Smith (Eds.), *Learner English* (2nd ed.,). Cambridge, England: Cambridge University Press. 279-295

Observations on the transitions from vowels to voiceless obstruents: a comparative study

Mechtild Tronnier
Centre for Languages and Literature, Lund University, Sweden

Abstract

It has been observed that the when segmenting running speech, placing a boundary between a vowel and a following voiceless stop is not always a trivial task. When discussing segmentation criteria in an international cooperation, it showed that different realisations of V+voiceless stop sequences occurred between the recorded material of the involved languages, which led to differences in where the boundary marker between the two segments was placed.

The study presented here aims to give an overview over which types of transitions that may occur in the phase from a vowel to a voiceless stop. Such transtions are related to the timing of two features, namely: oral closure for the stop and voice offset from the vowel. Data from speakers of Swedish, German and Italian has been analysed and quantity charactaristics of the different languages, vowel quality and place of articulation of the stop consonant has been taken into consideration as possible factors which promote different types of transitions.

Introduction

The transition from a vowel into a voiceless stop requires the coordination of oral closure with the offset of vocal fold vibration. In most cases, these two actions are aligned, but it may happen that the two processes occur at different points in time. In the case of a disalignment, a section similar to frication-like sound may turn up. In previous studies (Tronnier, 2002a, 2002b) such frication noise has been accounted for in a similar way as the phonological feature of preaspiration in Icelandic and some other Swedish dialects (cf. Helgasson, 2002).

In an ongoing project on the L2-acquisition of the contrast between singletons and geminates in Italian by L1-speakers of German on one hand and Swedish (Einfeldt et. al. 2017) on the other hand, the discussion about where to establish the end of the vowel in the case of a sequence of *V+voiceless obstruent* arose. This was due to the variation of alignment of the above mentioned processes. In most cases, Italian L1-speakers and German speakers of L2-Italian produced a proper alignment for oral closure and voice offset, and only occasionally a frication noise in the transition between a vowel and a voiceless stop. For that reason, voice offset was chosen as the criterion for the segment boundary inbetween and the frication noise was accounted for as part of the occlusive phase of the following stop or in case of a following voiceless fricative, as a part of that consonant. For the Swedish speakers of L2-Italian, however, large sections of frication noise in addition to remaining formant structure from the previous vowel was observed. This led to the choice to group the corresponding section to the vowel.

The following study aims to give a picture on how the transition between the elements of a sequence *V+voiceless stop* is carried out by L1-speakers of the three languages. Hereby, the type of realistion of the transition – i.e. with or without frication noise or even the occurrence of other glottal features – is explored with regard to different factors. These factors are:

1. The speakers' L1
2. The quantity characteristics in which the sequence takes place
3. Vowel quality
4. Place of articulation of the stop.

The study

Carrier sentences with target words containing the sequence of interest in German, Italian and Swedish were recorded, when read by L1-speakers. For all three languages, the speakers were asked to produce the target words – which were highlighted by bold print in the presented text – as a word in focus.

The list of focus words for all three languages contains samples with on the one

hand contrasting vowels /i/ and /a/ and on the other hand contrasting place of articulation for the consonant, namely bilabial and alveolar, resulting in /p/ and /t/. In addition, quantity contrast was included in the samples according to the quantity characteristics for each language. In that way, geminated and singleton consonants with otherwise similar traits in the sequence occurred for the target words in Italian. For German and Swedish, the contrast of long vs, short vowels was included in the choice of target words. For Swedish, such contrast may co-occur with complementary variation in consonant length.

For example:

Italian: *fata – fatta* [fata] – [fat:a]
German: *Rate – Ratte* [ra:tə] – [ratə]
Swedish: *mata – matta* [mɑ:ta] – [mat:a]

So far, eight speakers of each language have been recorded. They were randomnly chosen and are in all groups between 25 and 45 years of age.

The sound quality of the recordings varies, as the recording environment was chosen for practicality reasons. Some recordings had to be excluded from further analysis due to considerable echo-effects.

For the analysis of the data, the vowel, the stop and in the case that some transitional phase appeared in the target word, that section was manually segmented in Praat.

Preliminary results

The analysis of the recordings so far have shown that different types of transitions do occur, however with a varied degree. Most frequently an alignment of voice offset and oral closure is observable. Next to it in frequency, a pattern with an early devoicing is produced, which is followed by a frication noise before the oral closure sets in. This is the type of devoicing that has been observed in the earlier studies on L2-production (Einfeldt et. al. 2017). Other – rather rare – types that can be found in the data, consists of either a short phase of creaky voice or a phase of a clearly breathy vowel between a modal vowel and the onset of the occlusive phase of the following stop.

Further examination on whether other factors (cf. above, 1. – 4.) are influential on the appearance of a fricative transition, reveals so far only that language background is a convincing factor. Other factors, like a certain vowel quality, a certain consonantal place of articulation, quantity character do not seem to play a larger role. In that way, most of the speakers of Swedish produce such transitions in all conditions. The speakers of German and Italian produce that type of transition only occasionally and in case it does, no contextual influence seems to be the reason for it either.

The preliminary results from this rather exploratory study raise the question on whether fricative-transitions in Swedish are part of the sound characteristics of the language, however not being normative, but accepted.

References

Einfeldt, M., Kupisch, T. and Tronnier, M. (2017). The acquisition of Italian geminates by L1 German and Swedish speakers in production and perception. *Book of Abstracts: International Symposium on Monolingual and Bilingual Speech 2017*, Greece, Chania, 68-68.

Helgason, Pétur (2002). *Preaspiration in the Nordic Languages.* Stockholm University.

Tronnier, M. (2002). Preaspiration in Southern Swedish Dialects. *Proceedings of Fonetik 2002. Speech, Music and Hearing Quarterly Progress and Status Report*, 44: 33-36, KTH, Stockholm.

Tronnier, M. (2002). Preaspirated Stops in Southern Swedish. *Proceedings of ICSLP 2002*, Denver, Colorado.

Speech synthesis and evaluation at MTM

Christina Tånnander
Swedish Agency for Accessible Media

Abstract

The Swedish Agency for Accessible Media, MTM, is a government agency that provides people with reading difficulties with literature in accessible formats such as Braille and talking books. MTM uses speech synthesis to a large extent, mainly for producing talking newspapers and university textbooks.

This paper provides a historical account of how speech synthesis has been used and evaluated at MTM during the last decade, and discusses how MTM can develop more refined evaluation methods to ensure end users' best interests.

Introduction

This paper provides a historical account of how speech synthesis has been used at The Swedish Agency for Accessible Media (MTM) in the last decade, along with an overview of the user surveys and evaluation experiments performed; from the early acceptance tests from a time when MTM needed to know whether users could accept listening to synthetic speech, to Audience Response System-based tests to assure the quality of modern synthetic voices.

MTM and the Filibuster text-to-speech system

MTM, under the administration of the Ministry of Culture, produces and distributes literature, newspapers and periodicals in accessible formats, such as Braille, talking books and easy-to-read books, to persons with reading impairments. The agency was established in 1980 as the *Swedish Library of Talking Books and Braille* (TPB), but changed name to MTM in 2013.

MTM started using speech synthesis for Swedish and English university textbooks in the mid 00's. The Swedish voices at that time were not suited for university literature, wherefore MTM took the decision to develop their own text-to-speech (TTS) system, the unit selection system *Filibuster*. The Swedish male voice *Folke* was taken into production in 2007 (Ericsson et al., 2007; Sjölander et al., 2008). A commercial voice was used for English university textbooks. In 2009, the Norwegian Bokmål voice *Brage* was developed for the Norwegian counterpart to MTM, the Norwegian Library of Talking Books and Braille (NLB), (Sjölander & Tånnander, 2009), and in 2011 a female Swedish voice, *Tora*, was added to the Filibuster family. Finally, the Danish voice *Martin* was developed for Nota, the Danish counterpart to MTM in 2012.

Since 2017, MTM use commercial voices for both Swedish and English.

Speech synthesis evaluation

Methods for speech synthesis evaluation is a neglected area and there's a lack of knowledge about how to evaluate modern speech synthesis voices, especially for listening to long texts. Many evaluations do not consider the *ecological validity* of the evaluation method, that is to what extent the testing situation resembles the end-users' normal listening situation.

The well-known Blizzard Challenge, where speech synthesis developers build voices from the same speech database in order to compare the research methods, acknowledges the same problem (King, 2014). The organizers have called for better listening test designs and proposals of what to test, but few have responded. King points out two difficulties when deciding how to evaluate synthetic speech: "*First, it is not clear exactly what properties to evaluate. Second, it is hard to know how to evaluate the chosen properties, and one can never be certain that all of the listeners have correctly performed the task you expected of them.*" The quote pinpoints the complexity of the task, and speaks to the need for robust, carefully designed and performed evaluation methods.

MTM and TTS evaluation

Early acceptance tests

In the mid 00's, MTM performed three acceptance tests to find out whether the users could accept listening to university textbooks with speech synthesis.

Acceptance test I

One of the first listening tests of synthetic speech performed in collaboration with TPB was a master thesis using four existing Swedish TTS voices to explore if it was acceptable to use synthetic voices for university textbooks, and which enhancements that were required to make the voices good enough for reading university books (Persson, 2004). Eight subjects listened to the synthetic voices and one human voice. The results gave no straightforward answer to whether it's acceptable to use synthetic speech for university books.

Acceptance test II

A second acceptance test was performed in 2005, when TPB wanted to find out if the students using university talking books with human speech (usually without text) were ready for synthetic talking books with text (TPB, 2005). 12 subjects with dyslexia listened to a talking book, and answered questions about their experience. The response was scattered, but most students agreed that the reading experience was different from reading a talking book with a human voice.

When asked about listening fatigue, that is the phenomenon where the listener gets tired due to increased listening effort, six informants said that they could read for a shorter time than usual, while five meant that the length of their reading session wasn't affected.

Finally, a majority (nine students) answered that they could accept reading more synthetic talking books with text, though some of them under the condition that the voice must get better.

Acceptance test III

A similar survey was performed in 2006 (TPB, 2006). The target group were students that had borrowed a synthetic university talking book during the last months. 108 of 147 possible subjects took part in the survey, resulting in a response rate of 73%. 74% had dyslexia, 18% low vision, 5% were physically disabled and 3% had difficulties reading printed text for some other reason. 99% of the subjects had experience of talking books with human voices, and 58% had read books with synthetic speech.

47% students had read at least parts of the book. The reason for not having read the book was primarily that they had problems installing the reading system. More than 50% were satisfied with the synthetic book. Among the students that had read the whole book, 68% were satisfied, indicating either an increased tolerance over time, or that students that were satisfied from the beginning continued reading to the end. 82% thought the reading was effortful, and 38% that their comprehension was affected negatively.

Finally, 46% reported that they were open to borrow more synthetic books (the corresponding number for the students that thought it was easy to install the reading system was 61%). Yet 29% said that they could imagine borrowing synthetic books if they get better.

Synthetic speech for university textbooks

The first Swedish Filibuster voice

The first Swedish Filibuster voice, *Folke*, was evaluated in 2006, when it was compared to an existing, Swedish commercial voice as part of the quality assurance procedure. The evaluation consisted of different types of traditional methods such as grading of the general experience of the speech, discrimination tests (repeating words and names), and intelligibility ratings.

Comprehension test

In 2009, a master student performed a comprehension test of six texts of various length and complexity, read by the Swedish MTM synthetic voice *Folke* and the same texts read by the human voice behind *Folke* (Ståhl, 2009). 46 subjects, of which 19 were visually impaired, listened to the text, and answered questions about what they heard. The results show that the synthetic voice did not generate a noticeable impairment in comprehension compared to the human voice. The group with visually impaired subjects declared a higher self-experienced acceptance of the synthetic voice than the seeing group, but had a lower number of correct answers when listening to the synthetic voice (2.98 vs. 3.29). This difference was greatest in the longest and most complex text (1.70 vs. 3.00).

Student surveys

In a survey from 2009 examining how university students experienced downloading and streaming textbooks (TPB, 2009), the 19 students who had read a synthetic talking book during the test period were asked about their experience. 27% were positive (quite or very good), 21% neutral and 54% were negative (quite or very bad).

The next student survey was performed in 2010 (TPB, 2010), where 712 students were asked what they thought about MTM's Swedish and English university talking books produced

with synthetic speech. Their opinions about the books are shown in table 1.

Table 1. Opinion about synthetic talking books.

	Swedish	English
No opinion	24%	45%
Quite or very good	11%	10%
Neither good or bad	27%	15%
Quite or very bad	37%	30%

The user groups differed in how many that thought the books were quite or very good. The numbers are shown in table 2. The third column shows how many subjects per user group that were used to listening to synthetic speech.

Table 2. Quite/very good books and TTS experience per user group.

	Swe	Eng	Experience
Dyslexia	8%	7%	31%
Visually impaired	37%	62%	44%
Physically disabled	40%	20%	60%

Focus groups

In 2014, discussions with focus groups of 31 students with different reading disabilities were held (MTM, 2014). All of them had experience of talking books, both with human and synthetic speech. Most of them preferred a human voice, but for some students the text was so important that they preferred a synthetic talking book, where the text was attached. They pointed out that older human recordings can be of bad quality, and someone was annoyed of breathing sounds of a certain human reader. They thought that the female MTM voice Tora was one of the better Swedish voices. All participants think it's okay to get a book with a human voice with the register read by a synthetic voice.

Synthetic speech for newspapers

In 2010, MTM was instructed to perform a technological shift for talking newspapers (MTM, 2013b). Parts of the newspapers had earlier been recorded by human voices. The new talking newspaper model, where the entire newspaper was read by a commercial synthetic voice, allowed the subscribers to make their own selection of what to read or not to read. The newspapers were now digitally distributed to the end users' Daisy players or the app Legimus[1]. This technology shift also involved a user shift; university books with speech synthesis had been used primarily by the younger population, while the average age for a subscriber of a talking newspaper 79.3 years in 2013 (MTM, 2013a).

Field study

An evaluation of a field study where the users had tried reading newspapers with speech synthesis concerned mainly older people (MTM, 2012). Out of 54 subjects, only 11% were younger than 66 years old, and 63% were 76 years or older. 73% of them reported that they thought the synthetic voice was very or quite easy to understand, while 23% thought it was quite difficult or very difficult to understand and 5% didn't know. 52% considered the voice pleasant, 36% thought it was not so pleasant and 13% didn't know. They were also asked about which features they thought were most important for a synthetic voice, resulting in a ranking where the top three features were pronunciation (51%), followed by speed (32%), and humanlikeness (27%). Of the 53% that didn't read their paper on a regular basis, 75% reported that the problems were player-related. 28% said it took too long time, while 9% declared that the synthetic voice was the reason.

Evaluation of Daisy players

In a survey where 400 persons were interviewed about the Daisy player for reading newspapers in 2016 (MTM, 2016), 88% of the subjects reported that they thought the synthetic voice was clear. Still, they considered it boring, and thought that a more humanlike voice would make the reading experience more pleasant.

Subscriber survey

In 2017, 400 subscribers of talking newspapers were interviewed in order to find out what the they thought about reading the newspaper (MTM, 2017c). 78% thought that it was easy to listen to the paper and that the equipment worked well, and 73% found the synthetic voice clear and intelligible.

Terminated subscriptions

As a preparation for future procurements of reading equipment for talking newspapers, a survey that investigated why 200 newspaper

[1] https://www.mtm.se/produkter-och-tjanster/taltidningar/dagstidningen-som-taltidning/olika-satt-att-lasa-taltidningen/

http://www.legimus.se/appenlegimus

subscribers had terminated their subscription was carried out in 2017 (MTM, 2017d). The main reason turned out to be natural causes (62%) such as illness and age, followed by the equipment (15%) that was considered difficult to use. The synthetic voice was the reason for 5%, the newspaper itself for 4% and other reasons 5%.

Synthetic speech for fiction

English fiction – quantitative survey

In 2016, MTM wanted to find out if the users could accept reading English fiction literature with speech synthesis (MTM, 2017b). 350 subjects listened to an eight minutes sound file, where a synthetic voice read a passage from a fiction book. Most of the subjects weren't used to neither English or synthetic talking books, and only 2% had English as their first language.

43% had a good reading experience, 36% neither good or bad, 18% bad and 3% didn't know. 45% declared that they could accept listening to an entire book with this voice, 27% could not and 18% didn't know. Table 3 shows what the subjects answered to the questions 'what was good/bad with the reading?'.

Table 3. What was good/bad with the reading?

Feature	% good	% bad
Easy to understand	52	14
Pleasant voice	51	12
Flow, no clips	36	29
Words are pronounced correctly	43	13
Sounds like a human	39	26
Other	17	33
Dialogues read in a good way	n/a	34

66% thought that they can get an acceptable reading experience when they get used to the voice, and 46% meant that their reading experience could be the same as when listening to a human voice. On the contrary, 20% didn't think they could get used to the voice at all, and 27% that the reading voice couldn't be the same as for a human voice.

As for accepting synthetic talking books at all, 71% thought that it's better to get a talking book with this synthetic voice than not getting the book at all, while 16% meant that they'd rather have no book. 40% appreciated the idea of getting a synthetic talking book while waiting for a human recording of the book, but 44% would rather wait for the human recording. Finally, when asked about which genres they'd prefer listening to with speech synthesis, 65% answered non-fiction literature, 29% biographies, 29% novels, 26% detective stories or thrillers. Only 4% thought that poetry was a good genre for speech synthesis.

English fiction – qualitative survey

This quantitative survey was followed by qualitative telephone interviews of ten subjects (MTM, 2017a). They could choose one of nine books read by the English synthetic voice, and were instructed to listen to at least two hours. In general, they thought the voice was clear and intelligible, but considered it monotone. One of the most common problems was that it was difficult to know when a conversation starts and ends, and who's speaking. Some thought they got used to the errors of the voice, while others got more and more irritated. The results from the interviews thus showed scattered judgments about how it is to read an English fiction book with speech synthesis.

Web survey about voice preferences

In connection with a web survey about customer satisfaction, MTM took the opportunity to ask the subjects some questions about speech synthesis preferences, since MTM at the moment were working with a procurement of commercial speech synthesis voices for Swedish and English. When asked about which English (British or American) they preferred, 38% answered a British voice, 19% an American voice, and 43% answered that it didn't matter. 20% preferred a female voice, 15% a male voice, and 65% said it didn't matter. Next, the respondents were asked to choose one of four alternatives as the most and least important feature of a synthetic voice.

Table 4. Most and least important features of a synthetic voice.

Feature	Most	Least
Sounds like a human	40%	17%
Easy to understand	29%	8%
All words are pronounced correctly	21%	18%
Works well listening to at high speed	10%	57%

These results led to the procurement of a female Swedish voice and a male British voice, and showed that the subjects clearly valued humanlikeness and intelligibility over the other two feature choices.

Audience response system evaluations

Audio response system (ARS) is a method borrowed from the entertainment business, where the audience is instructed to click a button whenever they perceive a certain stimulus, for example something amusing, boring, or, for the evaluation of speech synthesis, something unintelligible or disturbing in the speech signal. The temporal precision and reliability of the ARS tests have been evaluated and proven robust in Edlund et al., 2013a, 2013b.

A novel experiment

In 2012, an ARS evaluation was tried out to better mirror the reading situation of a real end-user (Tånnander, 2012). This novel experiment compared three different instructions of when to click the button: click when you hear something (1) unintelligible, (2) irritating, and (3) not entirely correct. Three groups of eight subjects each listened to four sound files with synthetic speech, leading to the results in table 5, where we can see that group 1 clicked about ten times less than group 2 and 3.

Table 5. Total number of clicks and average number of clicks per file and subject.

Group	Total	Average/file and subject
1	28	0.88
2	272	8.50
3	294	9.18

Furthermore, the variation between the number of clicks varied a lot within group 2 and 3. The results also showed that the subjects agreed upon where the synthetic speech sounded bad, with clear peaks at points where the synthetic voice had problems with for example foreign names.

Audience response system-based assessment for analysis-by-synthesis

In the next experiment, 20 subjects listened to a three minutes excerpt from a university textbook read by a synthetic voice, and clicked a button on a Microsoft Xbox 360 controller every time they perceived an error or something they disliked (Edlund et al., 2015). The subjects clicked between 29 and 50 times each, and yet again, the distribution analysis showed consistency in when the subjects had reacted.

A professional synthesis developer analysed the 31 tallest clicking peaks, and found 68% of the peaks to be connected to an easily identified problem. 13% could not be identified and for 19%, no connected problem was found. Over the 14 tallest peaks, 100% could be connected to a problem, and over the 21 tallest peaks, 95% could be identified as belonging to a certain event in the synthetic speech. The authors summarised the results like this: *"(1) clicks generated by subjects are distributed such that clear peaks can be easily found; (2) peaks correlate with known internal states associated with quality; (3) we can find the average response latency; (4) in most of cases, a professional speech synthesis developer can find what likely caused subjects to click, (5) especially for high peaks."*

Other ARS evaluations

Furthermore, the ARS evaluation method has been used in experiments evaluating intelligibility in deviant child speech (Strömbergsson & Tånnander, 2013) and the ranking of the speech errors' severity (Strömbergsson et al., 2014). In addition, the ARS method has been used by MTM in internal comparisons of synthetic voices.

Discussion

The overview reflects that a wide range of different user surveys and evaluations of speech synthesis and talking books and newspaper read by a synthetic voice has been used. The results have been of advisory use for strategic decisions, as well as for comparisons and quality assurance of synthetic voices.

Accessible evaluation tools

Most of the early surveys include personal interviews, which are time-consuming and expensive. However, due to the lack of accessible evaluation tools, it has not been possible to perform large-scale, quantitative surveys where all MTM's target groups can participate.

Holistic approaches and ecological validity

It is important to keep in mind that some of the above-mentioned surveys do not evaluate synthetic speech alone, but take a rather more holistic approach to reading synthetic talking books or newspapers, than a survey evaluating synthetic speech alone. For example, in acceptance test II (TPB, 2005), the acceptance rate of synthetic talking books decreased because the reading system was difficult to install, but increased because many subjects thought it was more important go get the text and a synthetic voice compared to a human voice without text.

Thus, the test concerned the total reading experience of a university talking book with visible text, read by a synthetic voice, and listened to with a certain reading system that had to be installed before the 'listening test' could begin. This is an example of a survey of high ecological validity, since installation and configuration of technical equipment or software is usually part of real-world reading preparations.

Future work

The ARS evaluation have proven to be a robust method which can point to temporally distinct events that are perceived in continuous, lengthy listening. MTM will continue to this type of evaluations in collaboration with TMH, KTH, as well as looking at which aspects of speech that give rise to listening fatigue. The same team are also working on methods to accurately measure comprehension and retainment of lengthy texts in for example learning situations, both as they relate to the voice/speech synthesis quality and as a function of other aspects of the reading system and the reading situation.

Finally, a deeper knowledge about the needs and preferences of various users and how these relate to different user characteristics would substantially strengthen MTM in its role as a modern knowledge centre.

References

Edlund, J, Al Moubayed, S, Tånnander, C, & Gustafson, J (2013a). Audience response system based evaluation of speech synthesis. In *Proc. of Fonetik 2013*. Linköping, Sweden.

Edlund, J, Al Moubayed, S, Tånnander, C, & Gustafson, J (2013b). Temporal precision and reliability of audience response system based annotation. In *Proc. of Multimodal Corpora 2013*. Glasgow, UK.

Edlund, J, Tånnander, C, & Gustafson, J (2015). Audience response system-based assessment for analysis-by-synthesis. In *Proc. of ICPhS*. Glasgow, UK.

Ericsson, C, Klein, J, Sjölander, K, & Sönnebo, L (2007). Filibuster – a new Swedish text-to-speech system. *Proc. of Fonetik 2007*.

King, S (2014). Measuring a decade of progress in Text-to-Speech ; Evaluando una década de avances en la conversión texto-habla. *Loquens*, *1*(1).

MTM (2012). *Utvärdering av fältförsöket med den nya taltidningsmodellen*.

MTM (2013a). *Användarundersökning T2*. Stockholm, Sweden.

MTM (2013b). *Taltidningen 2.0 slutrapport*. Stockholm, Sweden.

MTM (2014). *Fokusgrupper med studenter*. Stockholm, Sweden.

MTM (2016). *Utvärdering taltidningsspelare*. Stockholm, Sweden.

MTM (2017a). *Engelsk talsyntes: kvalitativ undersökning*. Stockholm, Sweden.

MTM (2017b). *Engelsk talsyntes: kvantitativ undersökning*. Stockholm, Sweden.

MTM (2017c). *Läsarundersökning*. Stockholm, Sweden.

MTM (2017d). *Uppsagda prenumeranter*. Stockholm, Sweden.

Persson, M (2004). *Utvärdering av några text-till-talomvandlare som läshjälpmedel*. Uppsala University.

Sjölander, K, Sönnebo, L, & Tånnander, C (2008). Recent advancements in the Filibuster text-to-speech system. In *Proc. of the Swedish Language Technology Conference (SLTC)*. Stockholm, Sweden.

Sjölander, K, & Tånnander, C (2009). Adapting the Filibuster text-to-speech system for Norwegian bokmål. In *Proceedings of Fonetik 2009*.

Strömbergsson, S, & Tånnander, C (2013). Correlates to intelligibility in deviant child speech - Comparing clinical evaluations to audience response system-based evaluations by untrained listeners. In *Proc. of Interspeech*. Lyon, France.

Strömbergsson, S, Tånnander, C, & Edlund, J (2014). Ranking severity of speech errors by their phonological impact in context. In *Proc. of Interspeech*. Singapore.

Ståhl, M (2009). *Folke vs Henry: En jämförelse av förståelse mellan syntetisk och mänsklig uppläsning av sammanhängande texter*. Stockholm University.

TPB (2005). *Acceptansundersökning II: Januari-februari 2005*. Stockholm, Sweden.

TPB (2006). *Acceptansundersökning III: En utvärdering av fulltextbok inläst med talsyntes till studenter*. Stockholm, Sweden.

TPB (2009). *Student Direkt: Uppföljning av testperioden*. Stockholm, Sweden.

TPB (2010). *Talboks- och punkskriftsbibliotekets studentenkät år 2010*. Stockholm, Sweden.

Tånnander, C (2012). An audience response system-based approach to speech synthesis evaluation. In *The Fourth Swedish Language Technology Conference (SLTC 2012)*. Lund, Sweden.

Teachers' opinion on the teaching of Swedish pronunciation

Elisabeth Zetterholm
Department of Language Education, Stockholm University

Abstract

In order to acquire more knowledge about teachers' opinions on their teaching of Swedish as a second language, a web-based survey was distributed. Questions about their teaching in general and more specific about pronunciation were asked. They reported that it is difficult to teach pronunciation since they need more training and pronunciation methodology in their formal education. However, they all mention that pronunciation instructions are of importance for second language learners of Swedish.

Introduction

For most second language learners, at least adults, a foreign accent is unavoidable (Moyer 2013). A native-like pronunciation is not necessary for an intelligible and comprehensible pronunciation and there is no clear correlation between foreign accent and intelligibility (Munro & Derwing 1995). It is shown that teaching instructions can improve learners' perception and production (e.g. Derwing & Munro 2015: Thomson & Derwing 2015). However, studies conducted in English-speaking countries on pronunciation teaching show that there is an underrepresentation of pronunciation instructions because teachers' do not receive enough training in their education (Breitkreutz, Derwing & Rossister 2001; Foote, Holtby & Derwing 2011). With this is mind, a similar study was conducted in Sweden (Zetterholm, 2017).

The aim of the study

The primary aim of the study was to determine to what extent pronunciation is taught in classroom at SFI (Swedish for Immigrants). Another purpose was to get more knowledge about teachers' overall goal when teaching Swedish as a second language.

Material and Method

A web-based survey containing questions about teachers' experience of teaching Swedish as a second language was distributed to teachers all over Sweden. 92 participants answered questions about their overall goal for teaching Swedish as a second language as well as questions with a specific focus on the teaching of pronunciation.

The survey included yes/no, multiple choice, as well as open-ended questions. For some of the questions participants were given the option to write an additional comment voluntarily.

Results

The answers indicate that the overall goal when teaching Swedish as a second language is communication, reading, writing, grammar and vocabulary. Teachers mention that a listener friendly and comprehensible pronunciation is of importance but usually not taught explicitly. The participants comment that pronunciation is of importance but too difficult to teach since they do not have enough education or teacher training. Prosody, such as phrase intonation, word stress and word accents as well as vowel quantity and quality are difficult both for teachers and learners, according to the comments.

Discussion

When, or if, the teachers give the learners some pronunciation instructions it is in combination with lessons about grammar or vocabulary. Explicit pronunciation lessons or feedback is rare. Quite a lot of the answers indicate that teachers want to have more knowledge about Swedish pronunciation and a methodology for teaching second language learners. They ask for more explicit instructions in their own education. The results of the survey will be further described and discussed in the presentation.

References

Breitkreutz, J.A., Derwing, T.M. & Rossister, M.J. (2001). Pronunciation Teacing Practices in Canada. *Tesl. Canada Journal*, Vol. 19. No 1.

Derwing, T.M. & Munro, M.J. (2015). *Pronunciation Fundamentals. Evidence-based Perspectives for L2 Teaching and Research*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Foote, J.F., Holtby, A.K. & Derwing, T.M. (2011). Survey of the Teaching of Pronunciation in Adult ESL Programs in Canada 2010. *Tesl. Canada Journal*, Vol 29. No 1.

Moyer, A. (2013). *Foreign Accent. The Phenomenon of Non-native Speech*. Cambridge University Press.

Munro, M.J. & Derwing, T.M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45(1): 73-97.

Thomson, R.I. & Derwing; T.M. (2015). The Effectiveness of L2 Pronunciation Instruction: A Narrative Revies. *Applied Linguistics* 2015: 36/3: 326-344.

Zetterholm, E. (2017). Swedish for immigrants. Teachers' opinion on the teaching of pronunciation. *Proceedings of the International Symposium on Monolingual and Bilingual Speech* 2017: 308-312.

Reduce speed now… for an intelligible pronunciation

Elisabeth Zetterholm[1], Harald Emgård[2] and Birgitta Vahlén[2]
[1]Department of Language Education, Stockholm University, [2]Självbildarna AB, Malmö

Abstract

This study investigates how explicit pronunciation instructions and training affect the speech of L2 learners, and if reduced articulation rate and overall rate improves comprehensibility. The participants, non-native speakers of Swedish with different first languages, received specific pronunciation training at their workplace twice a week for five weeks. The results show a reduced articulation rate as well as a more native-like rhythm and phrase intonation, and a significant progression in comprehensibility.

Introduction

Comprehensible and intelligible speech in a second language is crucial when learners need to integrate in the society and get a job. A non-native pronunciation can be difficult to modify, and the degree of foreign accent can have an effect on listeners' attitude (Abelin & Boyd, 2000; Munro, Derwing & Sato, 2006). Listeners' perception of L2 speech is often affected by speaking rates as well as segmental and prosodic errors in the utterance (Lennon, 1990; Munro & Derwing, 2001). L2 learners tend to speak slower than native speakers, and one reason might be articulatory difficulties that can cause repetitions and self-corrections. In order to improve the communication skills and comprehensibility, employees in the correctional system received explicit pronunciation instructions during a specific training program at the workplace. The training focused on prosody, e.g. articulation rate, fluency and intonation.

This is collaborative project between one researcher and two speech therapists teaching pronunciation in Swedish as a second language.

The aim of the study

The aim is to investigate if the pronunciation instructions and training have any effect on the learners' speech. Is the learners' speech more comprehensible if they reduce the articulation rate and capture a more native-like intonation?

Material and participants

Non-native speakers of Swedish attending a course in pronunciation at their workplace were recorded several times during the five week long course. This was done in order to document their progression of both read-aloud and spontaneous speech. These recordings were used for analyses in this study.

All participants are adults with different first languages and they have all completed studies in Swedish at SFI (Swedish for immigrants), but they need to improve their pronunciation to be able to carry out tasks and communicate in their work places.

Results

Assessments, made by three speech therapists, show a significant progression in verbal communication and comprehensability. Auditory and acoustic analyses confirms a reduced articulation rate, less repetitions and self-corrections, improvements in perceived fluency and a more native-like rhythm and phrase intonation.

Discussion

The results indicate that the rate of L2 learners' speech as well as perceived fluency, seems to be important factors for comprehensibility. The observations in this study confirm results from other studies, namely that explicit pronunciation instructions develop learners' speech, even short-time instructions (see Gordon & Darcy, 2016 for an overview)

References

Abelin, Å. & Boyd, S. (2000). Voice quality, foreign accent and attitudes to speakers. *Proceedings of Fonetik 2000*. Inst för Högskolan i Skövde; 21-24.

Gordon, J. & Darcy, I. (2016). The development of comprehensible speech in L2 learners. A classroom study on the effects of short-term pronunciation instruction. *Journal of Second Language Pronunciation* 2:1 (2016), 56-92. DOI 10.1075/jslp.2.1.03gor.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, 40: 387-417.

Munro, M.J. & Derwing, T.M. (2001). Modeling Perceptions of the Accentedness and Comprehensibility of L2 Speech The role of Speaking Rate. *Studies in Second Language Acquisition,* 23: 451-468.

Munro, M.J., Derwing, T.M. & Sato, K. (2006). Salient accents, covert attitudes: Consciousness-raising for pre-service second language teachers. *Prospect*, Vol. 21, No.1.